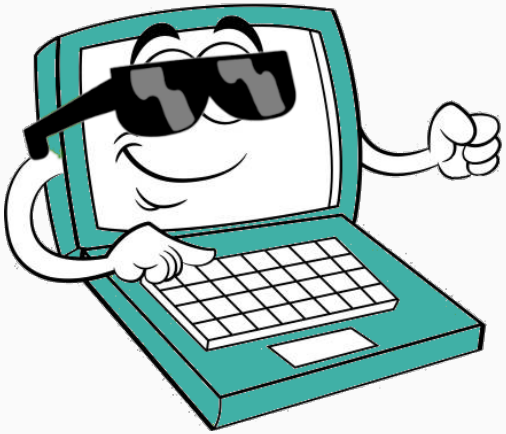
Methodology
Lesson 4: Evaluating the models

Carlos Ramisch
first.last@lis-lab.fr

M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

The Awesome system

My brand new Awesome system...



should soon replace

...the old Baseline system!

Expectation...

	dataset	metric1	metric2	metric3[1]
Baseline system	DS1	82.3	75.9	48.0
Awesome system	DS1	**95.3**	**89.8**	**65.4**
Baseline system	DS2	67.7	65.2	56.8
Aweseome system	DS2	**80.3**	**91.1**	**69.8**
Baseline system	DS3	77.6	74.1	92.8
Awesome system	DS3	**84.9**	**78.3**	**98.1**

$\Longrightarrow$ The Awesome system is better than the Baseline! 

[1] Higher is better

. . . Vs. reality!

	dataset	metric1	metric2	metric3
Baseline system	DS1	**82.3**	75.9	48.0
Awesome system	DS1	80.7	**76.2**	**50.4**
Baseline system	DS2	67.7	**65.2**	**56.8**
Awesome system	DS2	**67.9**	nan	49.6
Baseline system	DS3	77.6	**74.1**	92.8
Awesome system	DS3	**79.0**	**74.1**	**93.4**

$\Longrightarrow$ Wake up and smell the coffee 

Results analysis

- Identify overall trends
- Identify potential sources of problems (or bugs)
- Ensure conclusions are valid, claims are (statistically) sound

Experimental results

- Diversity of experiments $\Longrightarrow$ diversity of results
  - $\rightarrow$ Task at hand
  - $\rightarrow$ Datasets
  - $\rightarrow$ Evaluation metrics
  - $\rightarrow$ ...
- This course: no silver bullet, rather a toolbox

Outline

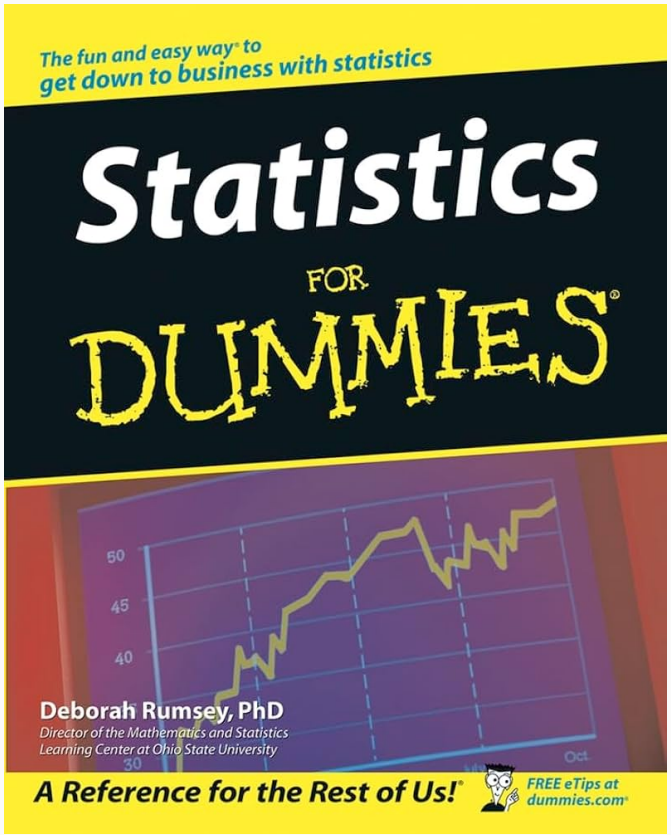
Statistics

- A mathematical framework to analyse data
  - $\longrightarrow$ Foundations: probability theory
- Statistical inference $\Longrightarrow$ data science, machine learning
  - $\longrightarrow$ Also: finances, health, biology, physics, social sciences, ...
- Identify trends, test hypotheses, measure correlations, ...

The problem with statistics

Finding good learning materials in statistics is hard

Too applied:



Too theoretical:

Weak Law of Large Numbers

The weak law of large numbers (cf. the strong law of large numbers) is a result in probability theory also known as Bernoulli's theorem. Let $X_1, ..., X_n$ be a sequence of independent and identically distributed random variables, each having a mean $\langle X_i \rangle = \mu$ and standard deviati σ. Define a new variable

$$X \equiv \frac{X_1 + \ldots + X_n}{n}.$$

Then, as $n \to \infty$, the sample mean $\langle x \rangle$ equals the population mean μ of each variable.

$$\begin{aligned}\langle X \rangle &= \left\langle \frac{X_1 + \ldots + X_n}{n} \right\rangle \\ &= \frac{1}{n} \left(\langle X_1 \rangle + \ldots + \langle X_n \rangle \right) \\ &= \frac{n\,\mu}{n} \\ &= \mu.\end{aligned}$$

In addition,

$$\begin{aligned}\operatorname{var}(X) &= \operatorname{var}\left(\frac{X_1 + \ldots + X_n}{n} \right) \\ &= \operatorname{var}\left(\frac{X_1}{n} \right) + \ldots + \operatorname{var}\left(\frac{X_n}{n} \right) \\ &= \frac{\sigma^2}{n^2} + \ldots + \frac{\sigma^2}{n^2} \\ &= \frac{\sigma^2}{n}.\end{aligned}$$

Therefore, by the Chebyshev inequality, for all $\epsilon > 0$,

$$P\left(|X - \mu| \geq \epsilon \right) \leq \frac{\operatorname{var}(X)}{\epsilon^2} = \frac{\sigma^2}{n\,\epsilon^2}.$$

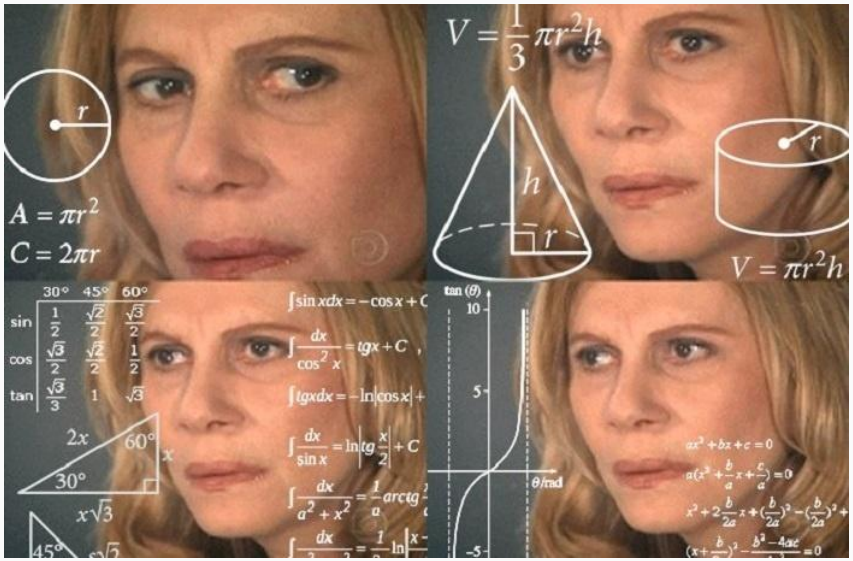
What usually happens

- A given statistical tool is used without (full) understanding
- Statistical tools applied because supervisor/reviewer asked
- Give up trying to understand, just use it as a blackbox

Truth be told: everyone hates statistics

Probability and statistics:

Difficult math, boring and totally useless, everyone hates it!



Truth be told: everyone hates statistics

Probability and statistics:

~~Difficult math, totally useless and so boring, everyone hates it!~~

- Difficult: mostly sums and products of fractions
- Boring: that's subjective, but yes, it may be boring
- Useless: definitely not! The basis of modern empiricism

Truth be told: everyone hates statistics

Probability and statistics:

~~Difficult math, totally useless and so boring, everyone hates it!~~

- Yes, we may hate it, but we also need it!
  - $\longrightarrow$ Knowing what we're doing can make us feel more at ease
  - $\longrightarrow$ It is worth the effort of overcoming initial resistance

Truth be told: everyone hates statistics

Probability and statistics:

- Yes, we may hate it, but we also need it!
  - $\longrightarrow$ Knowing what we're doing can make us feel more at ease
  - $\longrightarrow$ It is worth the effort of overcoming initial resistance

Ready? Let's go!

Random variable: definition

- A random variable is a variable with no specific value
  - $\longrightarrow$ It takes some value within a (known) set of possible values
  - $\longrightarrow$ We are not interested in its actual value

Examples:

- A human's age takes values form 0 to 130 years
- Water temperature at sea level ranges from 0°C to 100°C
- A person's handedness can be righ-handed, left-handed, both

Wooclap

Wooclap time!

Random variable: examples

Are the following (interesting) random variables?

- 1. The number of tentacles of an octopus?

Random variable: examples

Are the following (interesting) random variables?

- 1. The number of tentacles of an octopus?

  $\longrightarrow$ No, always the same value

Random variable: examples

Are the following (interesting) random variables?

- 2. The distance between the Earth and the Moon?

Random variable: examples

Are the following (interesting) random variables?

- 2. The distance between the Earth and the Moon?

  $\longrightarrow$ Yes, it actually varies from 363K to 406K km

Random variable: examples

Are the following (interesting) random variables?

- 3. A person's vote in the last presidential elections?

Random variable: examples

Are the following (interesting) random variables?

- 3. A person's vote in the last presidential elections?

  $\longrightarrow$ Yes, the values are the candidates/parties running

Random variable: examples

Are the following (interesting) random variables?

- 4. A person's opinion about how cute an octopus is?

Random variable: examples

Are the following (interesting) random variables?

- 4. A person's opinion about how cute an octopus is?
  - $\longrightarrow$ No, ill-defined, no closed set of possible values
  - $\longrightarrow$ Actually, everyone finds them cute! ;-)

Random variable: examples

Are the following (interesting) random variables?

- 1. The number of tentacles of an octopus? No
- 2. The distance between the Earth and the Moon? Yes
- 3. A person's vote in the last presidential elections? Yes
- 4. A person's opinion about how cute an octopus is? No

In short

- A variable is not random if its value is fixed / constant
- We need to be able to describe its set of possible values

  $\longrightarrow$ The set may be infinite (e.g. real numbers)

Why do we need random variables?

- Use their characteristics to understand the data
- Model features and evaluation metrics as random variables
- Basic block in probability and statistics
  - $\longrightarrow$ People have been studying them for a while
  - $\longrightarrow$ Statistical tools associated to them can be useful

Probability distributions

- Random variables come with probability distributions

$$P\{X = a\} = p(a) = 0.8 = 80\%$$

Probability distributions

- Random variables come with probability distributions

$$P\{X = a\} = p(a) = 0.8 = 80\%$$

- X: The random variable that we're interested in
- a: The particular value of that random variable
- 0.8 (80%): The probability that variable X takes value a

Probability distributions

- Random variables come with probability distributions

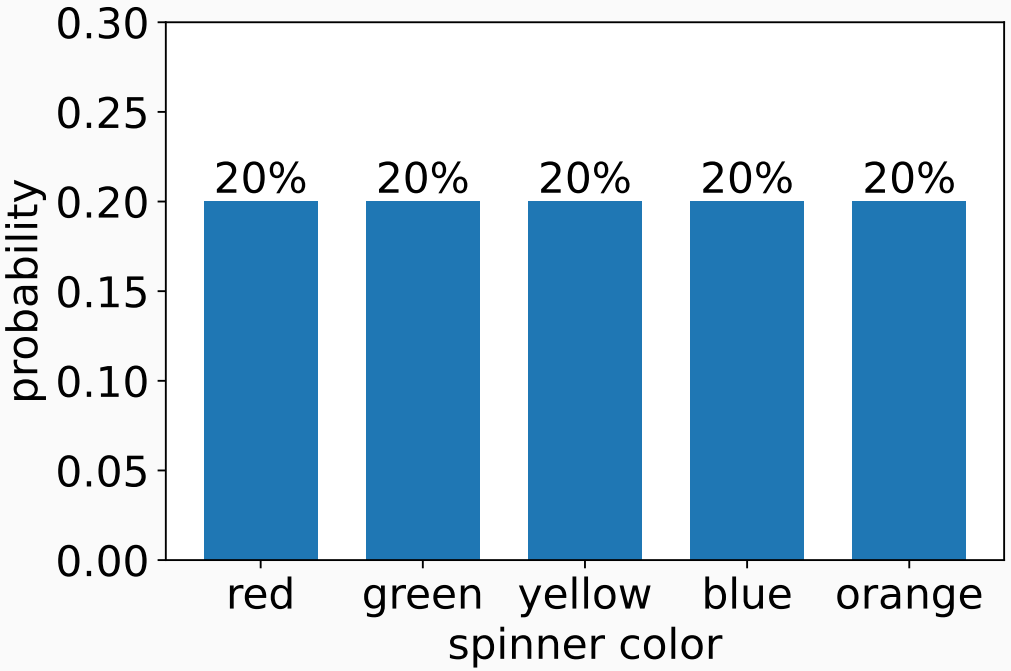
$$P\{X = a\} = p(a) = 0.8 = 80\%$$

- X: The random variable that we're interested in
- a: The particular value of that random variable
- 0.8 (80%): The probability that variable X takes value a

Note: probabilities $p(a)$ must sum up to 1 (for all values a)

Example probability distributions

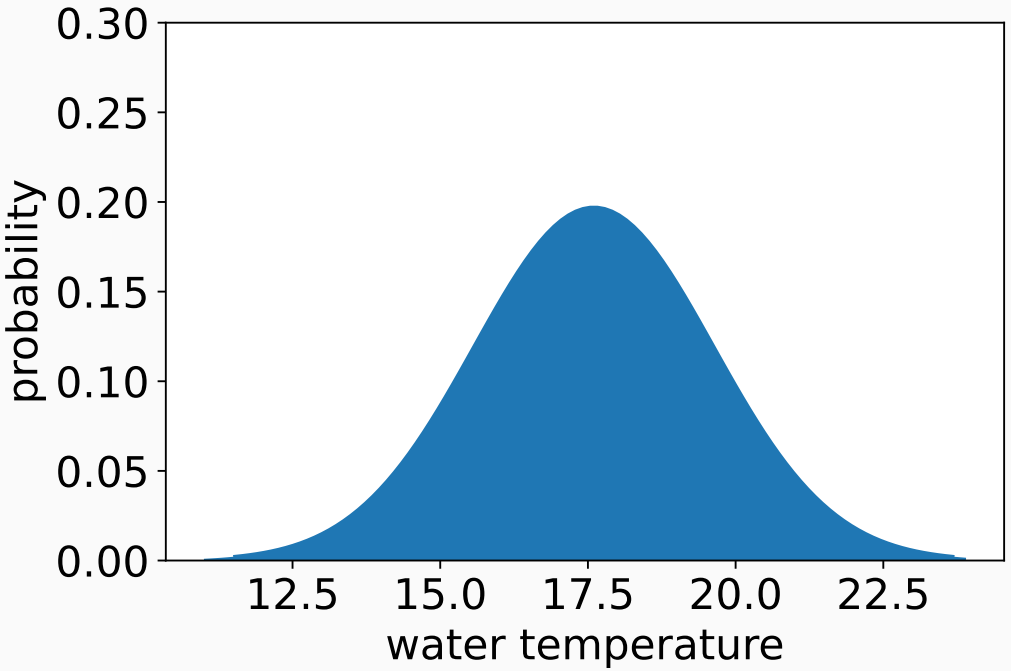
- X_1: color of a 5-coloured spinner wheel



$$P\{X_1 = \text{red}\} = p(\text{green}) = \ldots = p(\text{orange}) = \frac{1}{5}$$

Example probability distributions

- X_2: sea water temperature in July in Marseille



$$P\{X_3 < 17.6\} = 0.5$$

Wooclap

Wooclap time!

Probability distribution or not?

Which of the following are proper probability distributions? Why?

a)

x_i	$p(x_i)$
1	0.4
2	-0.2
3	0.8

b)

x_i	$p(x_i)$
0.4	0.4
0.35	0.35
0.25	0.25

c)

x_i	$p(x_i)$
-1	0.4
-2	0.2
-3	0.8

d)

x_i	$p(x_i)$
-1	0.4
0	0.2
1	0.2
2	0.1

Probability distribution or not?

Which of the following are proper probability distributions? Why?

a)

x_i	$p(x_i)$
1	0.4
2	-0.2
3	0.8

No, $p(2) < 0$

b)

x_i	$p(x_i)$
0.4	0.4
0.35	0.35
0.25	0.25

Yes, sum=1

c)

x_i	$p(x_i)$
-1	0.4
-2	0.2
-3	0.8

No, sum > 1

d)

x_i	$p(x_i)$
-1	0.4
0	0.2
1	0.2
2	0.1

No, sum < 1

From probabilities to statistics

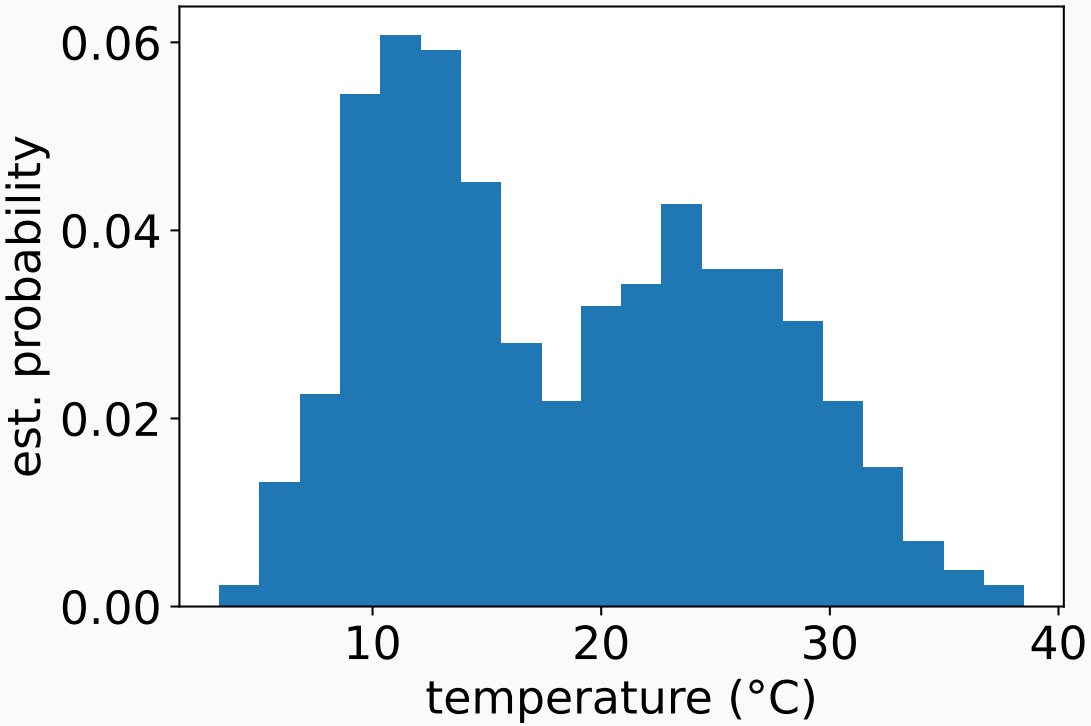
- Probability distributions are theoretical abstractions
  - $\rightarrow$ We often learn probabilities with toy examples
  - $\rightarrow$ In practice, X's "real" distribution is not accessible
- A sample is often used to estimate the probabilities
  - $\rightarrow$ Most of the time, probabilities are approximated
  - $\rightarrow$ Proportion in sample (%) $\rightarrow$ estimated probability

$$\frac{\text{count}(a)}{n} \approx P\{X = a\}$$

Random samples

- Randomly collect a finite set of data points
- **Example:** temperature of a sensor in a power plant

  → Size: 365 days → `[10.1, 14.0, 8.9,...12.5, 15.3, 13.3]`



Estimated probability distribution = normalized histogram

Sampling: example

Jupyter notebook 1 & 2

1. Open the dataset using `pandas.read_csv()`
2. Explore the different columns and their values
3. Make a histogram of the `compositionality` column

   → This is an estimate of its distribution!

Compositionality dataset

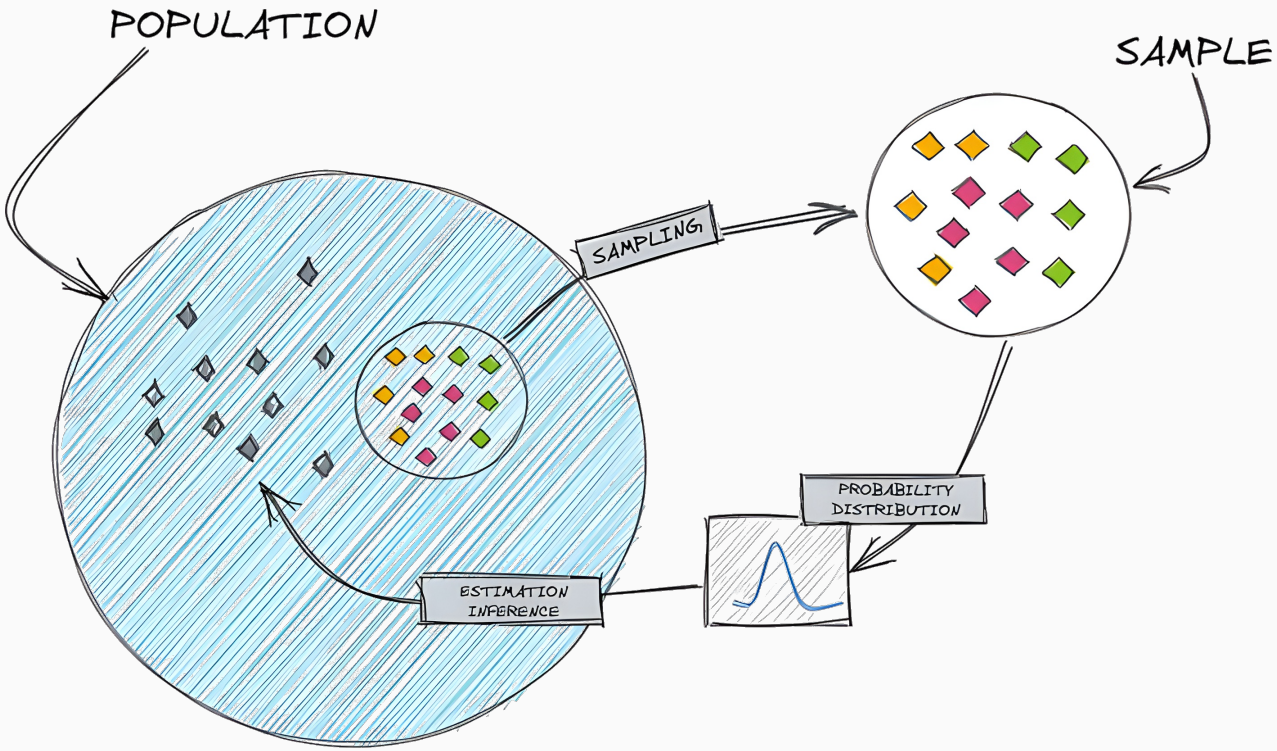
- *Is a dry run literally a run which is dry?*

  → not at all ←0 - 1 - 2 - 3 - 4 - 5 → absolutely yes
- Compositionality score: average rating of 10-15 annotators
- Sample: 180 compounds in French

Source: `https://aclanthology.org/J19-1001/`

Why do we need samples?

- A representative sample can inform us about the whole
  - $\rightarrow$ Full data not available: infer properties of (unknown) distribution
  - $\rightarrow$ Draw generalisable conclusions in the presence of uncertainty

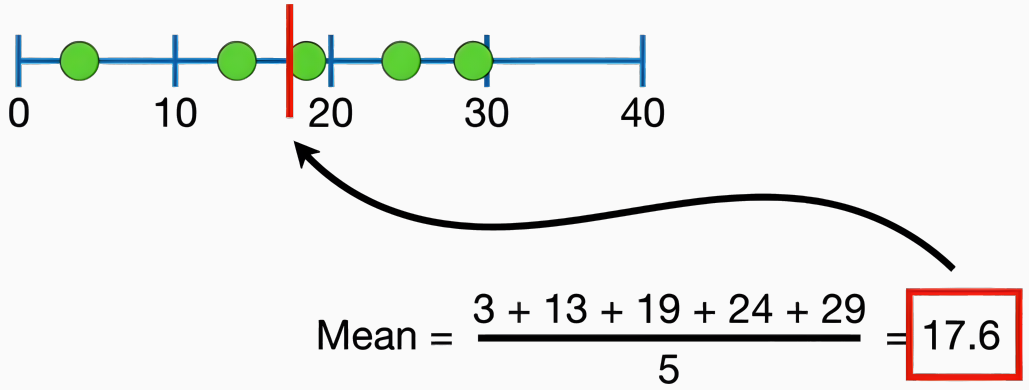


Sample mean / average

- A single value at the center of the sample

  $\rightarrow$ Summarise the whole data set with a single number

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i$$



Wooclap

Wooclap time!

Mean / average quiz

- Is the mean a probability (value between 0 and 1)?

Mean / average quiz

- Is the mean a probability (value between 0 and 1)?

  $\longrightarrow$ No, it depends on the values (arbitrary range)

- Is the value of the mean contained in the sample?

Mean / average quiz

- Is the mean a probability (value between 0 and 1)?

  $\rightarrow$ No, it depends on the values (arbitrary range)

- Is the value of the mean contained in the sample?

  $\rightarrow$ No, it can be a new value, not contained in the sample

- Is the value of the mean always positive?

Mean / average quiz

- Is the mean a probability (value between 0 and 1)?

  $\longrightarrow$ No, it depends on the values (arbitrary range)

- Is the value of the mean contained in the sample?

  $\longrightarrow$ No, it can be a new value, not contained in the sample

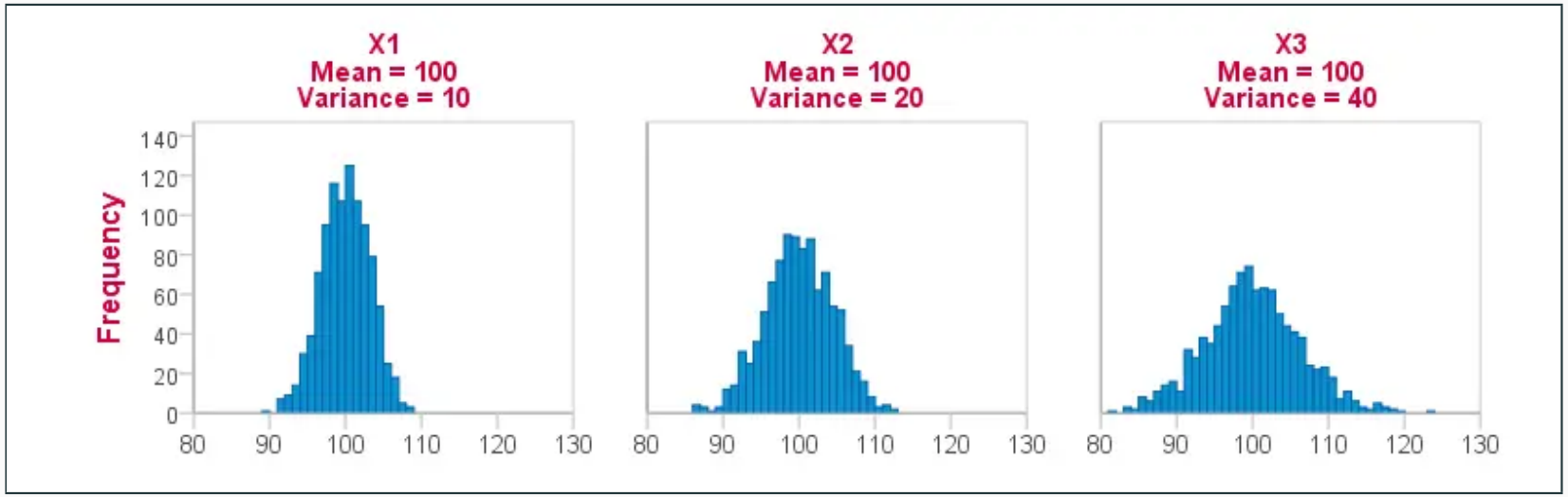
- Is the value of the mean always positive?

  $\longrightarrow$ No, e.g. if the variable only takes negative values

Data dispersion

- Mean does not take into account data dispersion

$$
\begin{array}{lcl}
S_1 & [0, 0, 0, 0] & \overline{S_1} = 0 \\
S_2 & [-4, -4, 4, 4] & \overline{S_2} = 0 \\
S_3 & [-6, -2, 1, 7] & \overline{S_3} = 0 \\
S_4 & [-15000, 15000] & \overline{S_4} = 0
\end{array}
$$



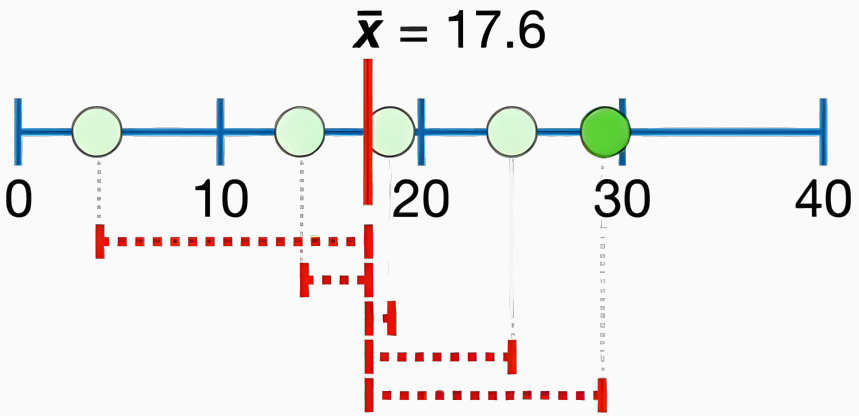
Source: https://www.spss-tutorials.com/descriptive-statistics-one-metric-variable/

Variance

- Variance characterises the dispersion/spread of a distribution
  - $\rightarrow$ Intuition: average distance from the mean
  - $\rightarrow$ $(x_i - \overline{x})$ can be positive or negative $\implies$ square it!

$$Var(X) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}$$

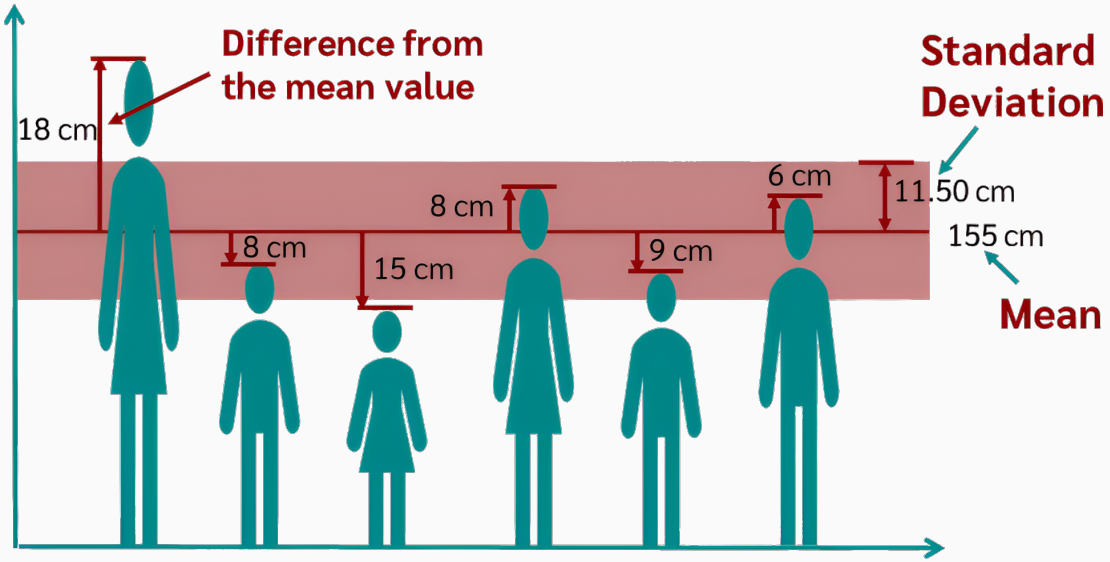
  - $\rightarrow$ Variance is always positive, differently from mean



Standard deviation

- Variance averages *squared* differences
  - $\rightarrow$ Its absolute value is hard to interpret
  - $\rightarrow$ Bring back to original value range $\rightarrow$ squared root
- The squared root of variance is called standard deviation

$$\sigma = \sqrt{Var(X)}$$



Estimated standard deviation

- Population standard deviation:

$$\sigma_X = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}}$$

- Sample standard deviation, unbiased estimator:

$$s_X = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}}$$

- Why? `https://www.youtube.com/watch?v=sHRBg6BhKjI`

In practice, we only need s_X $\rightarrow$ Most stats libraries' default

Calculating mean and standard deviation

Jupyter notebook 3

1. Open dataset containing 180 compositionality scores
2. Use Pandas' `comp.describe()` to obtain a summary
3. Is the obtained standard deviation σ_X or s_X?

One distribution to rule them all

The Normal distribution

$$P\{a < X < b\} = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} exp^{\frac{x-\mu}{\sigma}}$$

One distribution to rule them all

The Normal distribution

$$P\{a < X < b\} = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} exp^{\frac{x-\mu}{\sigma}}$$

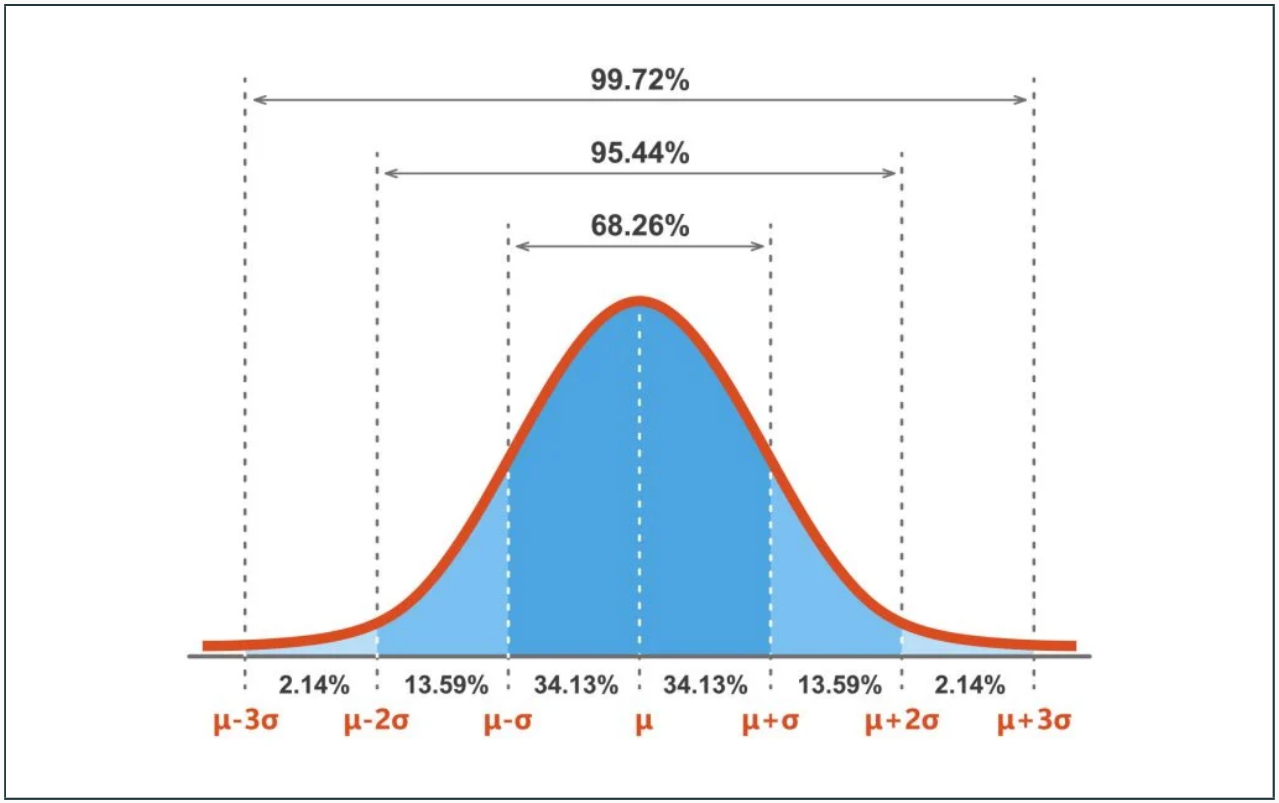
Who cares!

One distribution to rule them all

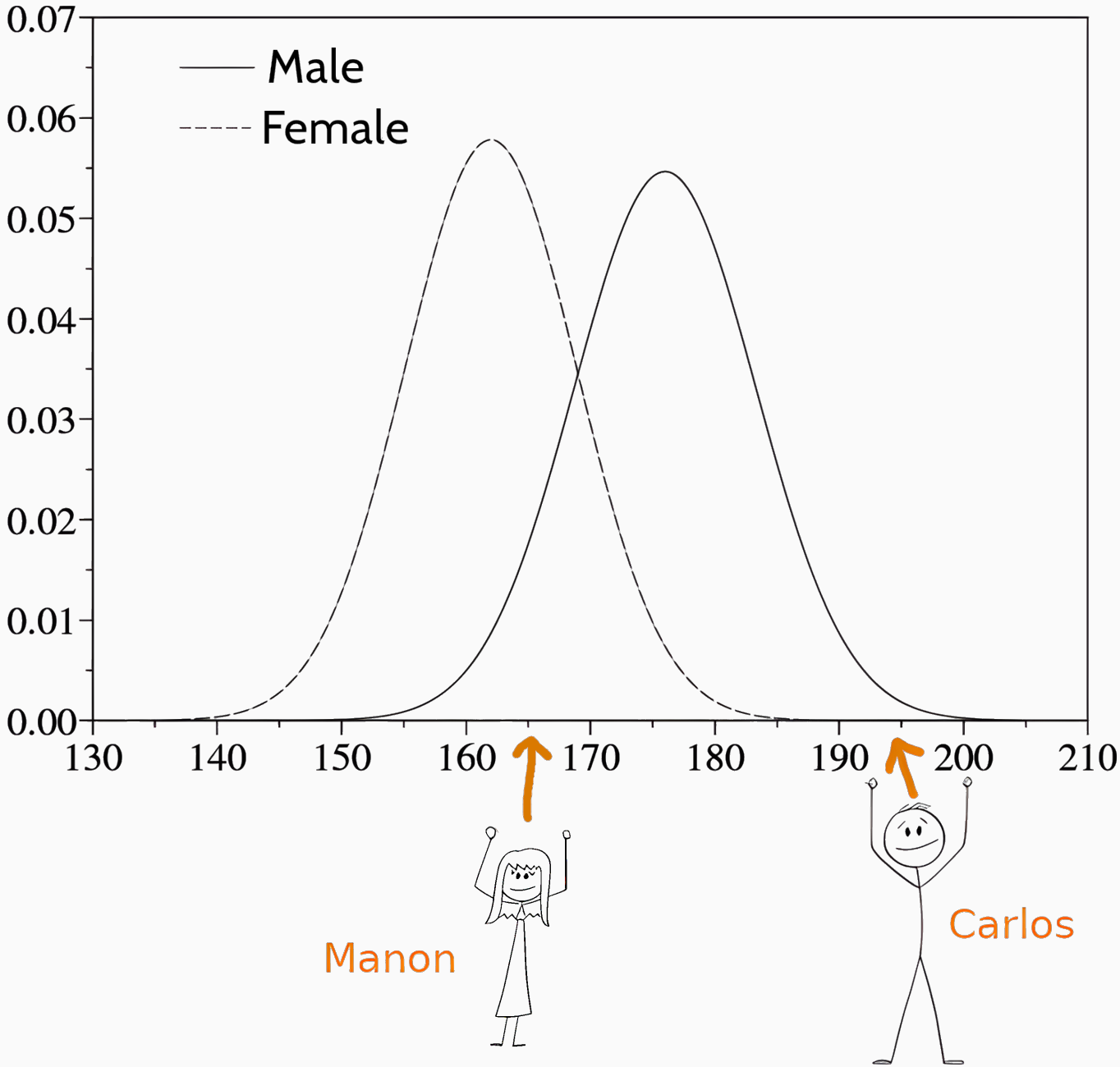
The Normal distribution

- Probability density function is a symmetric bell-shaped curve
- Specified by mean μ (center) and std. deviation σ (wideness)

  $\rightarrow$ 99.7% of probability between $\mu - 3\sigma$ and $\mu + 3\sigma$



Normal distribution: example

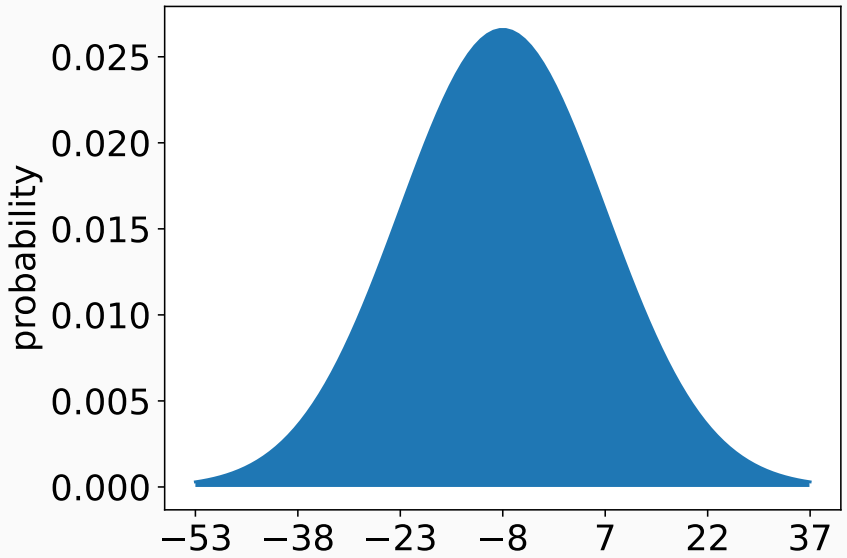


Wooclap

Wooclap time!

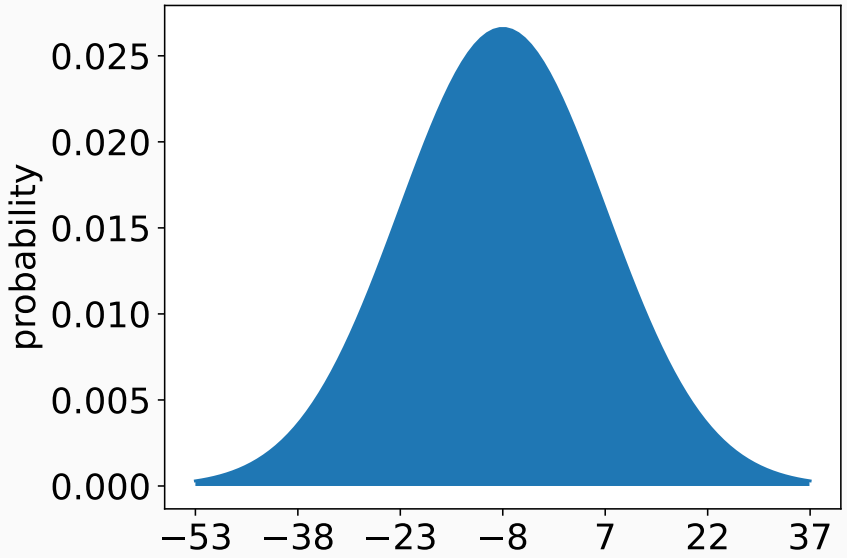
Who's that normal?

1. What are the μ and σ parameters for the following curve?



Who's that normal?

1. What are the μ and σ parameters for the following curve?

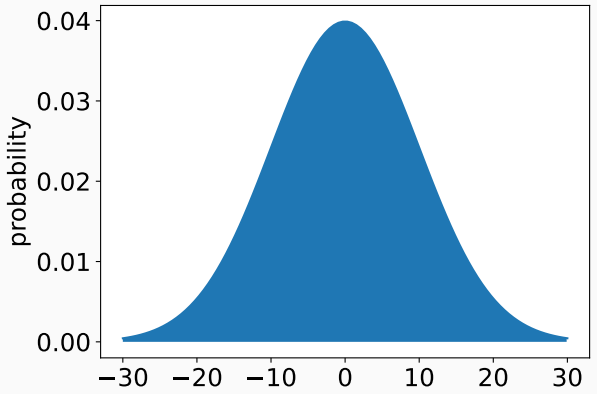


$\mu = -8$ and $\sigma = 15$

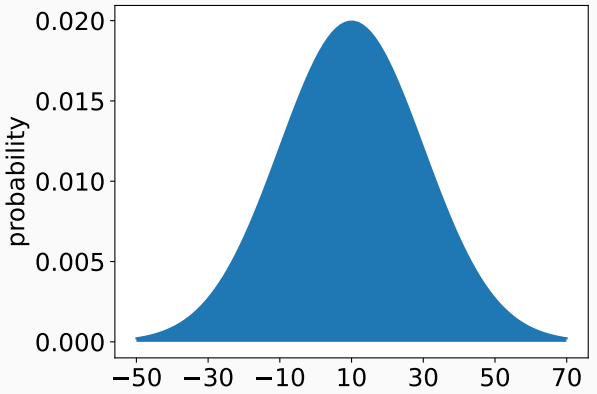
Who's that normal?

1. What are the μ and σ parameters for the following curve?
2. Which curve corresponds to $\mu = 10$ and $\sigma = 20$?

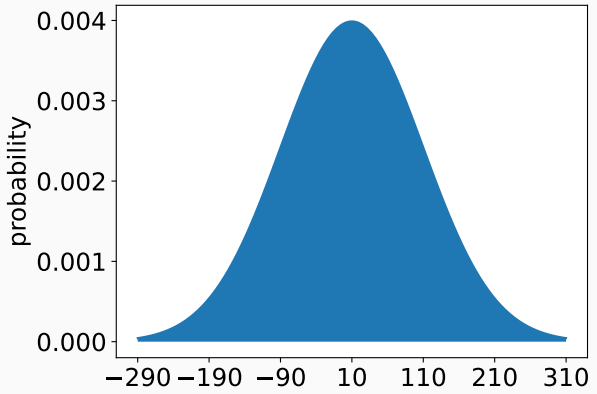
a)



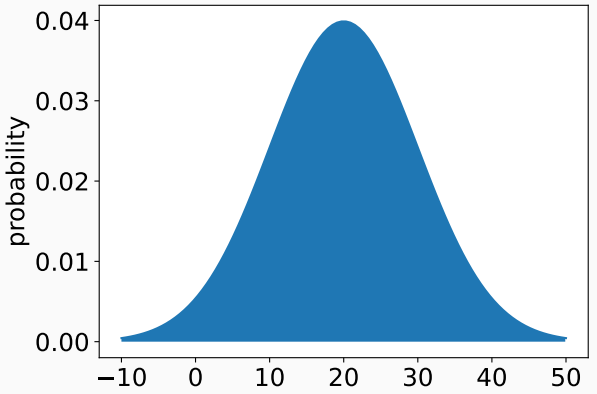
b)



c)



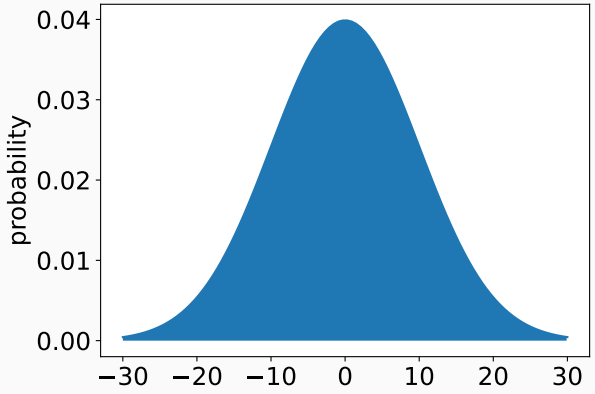
d)



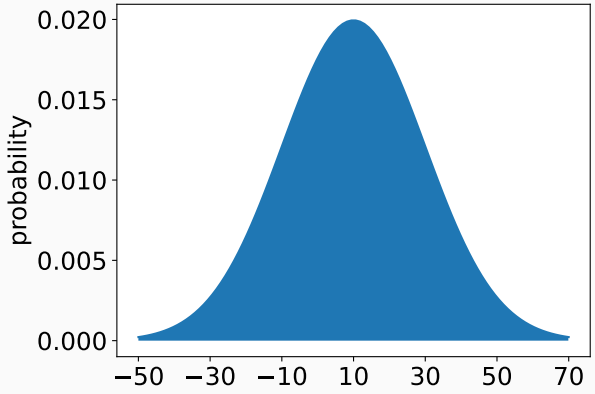
Who's that normal?

1. What are the μ and σ parameters for the following curve?
2. Which curve corresponds to $\mu = 10$ and $\sigma = 20$?

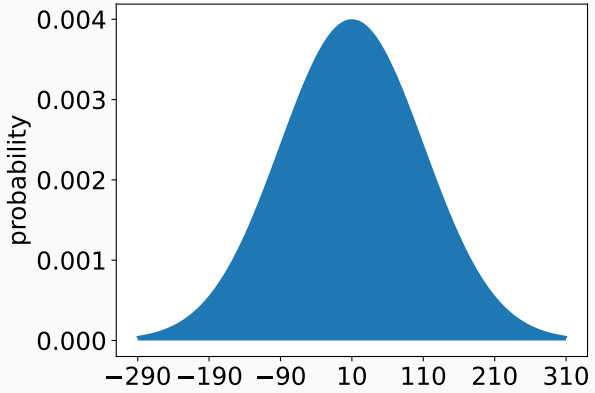
a)



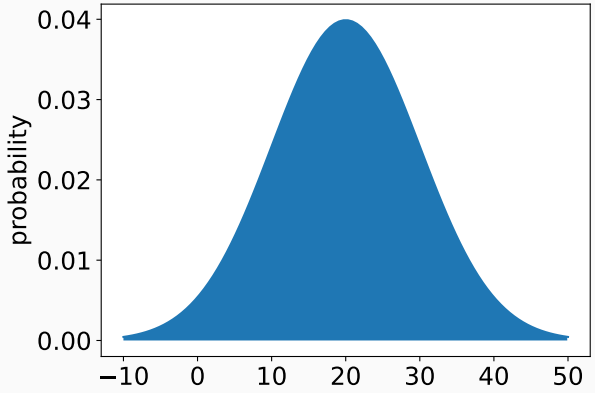
b)



c)



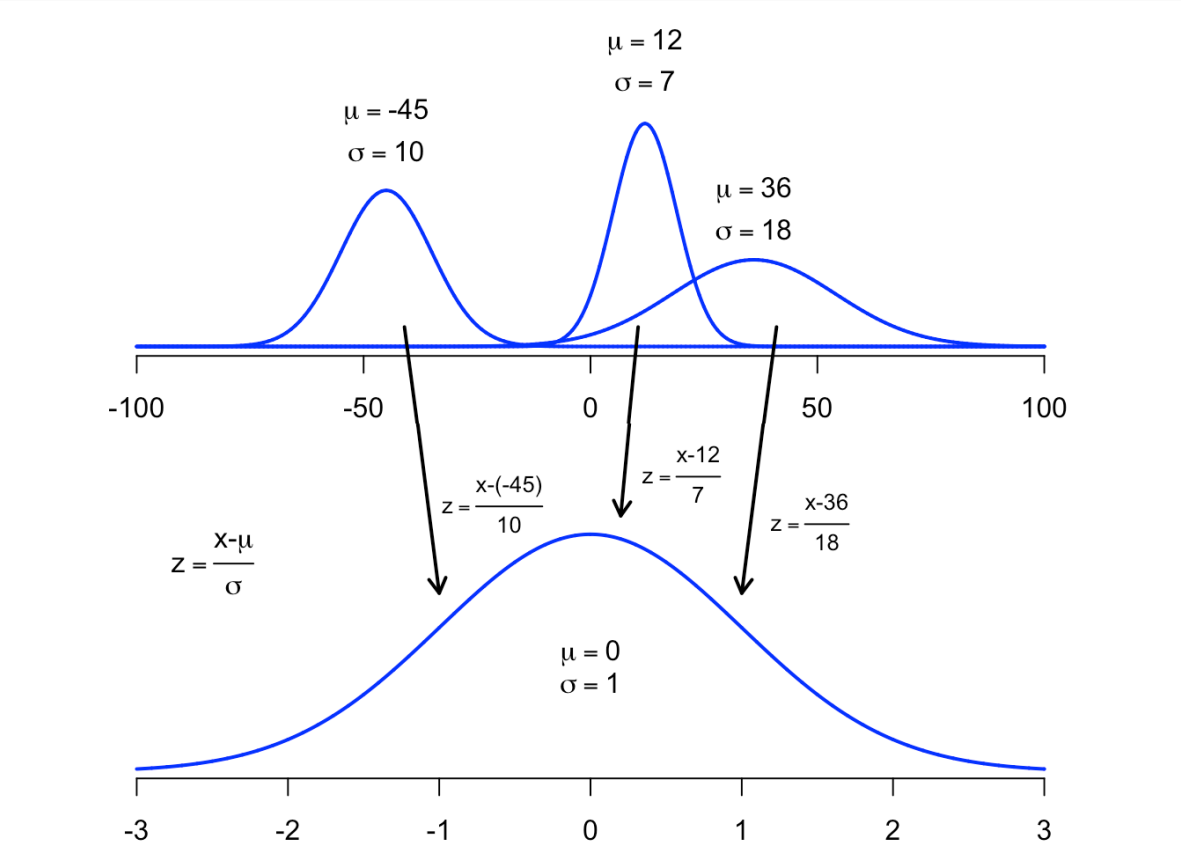
d)



curve b) – notice different heights

Standardization

- Calculate probability $\rightarrow$ integration (<o> aaaaah!)
  - $\rightarrow$ Normal is impossible to integrate analytically
- In practice:
  - $\rightarrow$ Standardize $z = \frac{x-\mu}{\sigma}$, then lookup table of $\Phi(a)$



Wooclap

Wooclap time!

The most famous probability distribution

Why is the normal distribution so important?

The most famous probability distribution

Why is the normal distribution so important?

- Turns out most measurements are normally distributed
- Used in many statistical tools, e.g. hypothesis testing
- Plays a central role in describing estimated means

It's normal to be average

- Normalised sum of random variables is normally distributed[2]
  - $\rightarrow$ Even if the variables are not normally distributed!
- The mean $\overline{x}$ of a sample is normally distributed
  - $\rightarrow$ Comes in handy to analyse averages
- This is known as the central limit theorem (CLT)

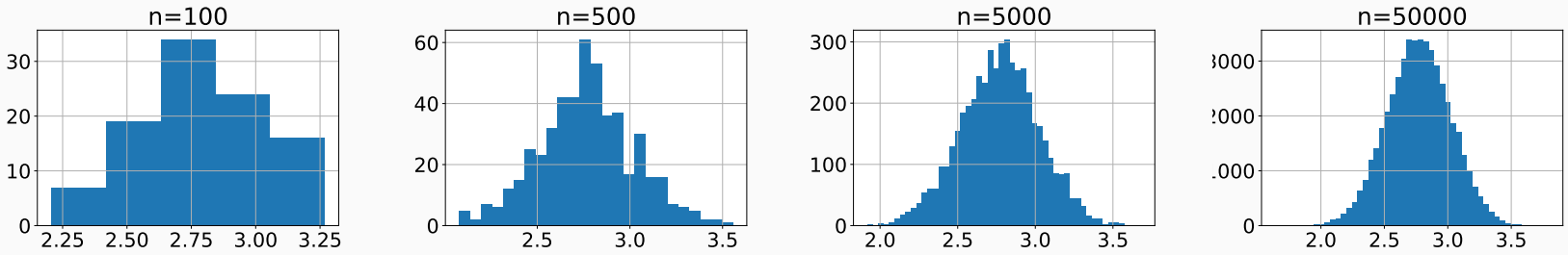
[2] Actually, independent and identically distributed (i.i.d.) random variables.

Central limit theorem: example

Jupyter notebook 4 & 5

1. Build n random samples of size 30 from compositionality data
2. Calculate mean of each random sample, save values
3. Estimate sample mean's distribution with histogram

   $\longrightarrow$ What happens when n increases?



In short

- Random variables and probability distributions
  - → Theoretical model for features and metrics
  - → In practice, estimated using sampling
- Mean and standard deviation characterise the data
- Normal distribution: bell shaped around the mean
  - → Useful to characterise values that are means (CLT)

In short

- Random variables and probability distributions
  - → Theoretical model for features and metrics
  - → In practice, estimated using sampling
- Mean and standard deviation characterise the data
- Normal distribution: bell shaped around the mean
  - → Useful to characterise values that are means (CLT)

Now we're ready for the next steps!

Outline

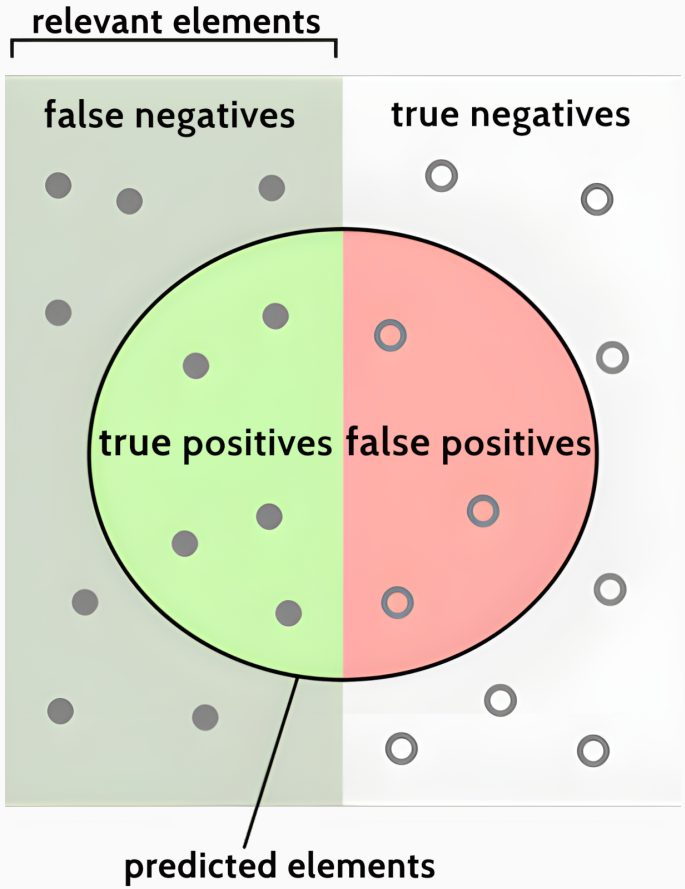
Disclaimer: all metrics are incomplete

- Ideally: measure a hidden variable or phenomenon
- In practice: measure what we can observe

  $\rightarrow$ Formulation is simple enough to be interpretable
- Metrics are partial views of the results

Classification framework

- *tp*: true positives
  → Correctly predict as positive
- *tn*: true negatives
  → Correctly predict as negative
- *fp*: false positives
  → Predict positive, should be negative
- *fn*: false negatives
  → Predict negative, should be positive



Source: image adapted from Wikipedia

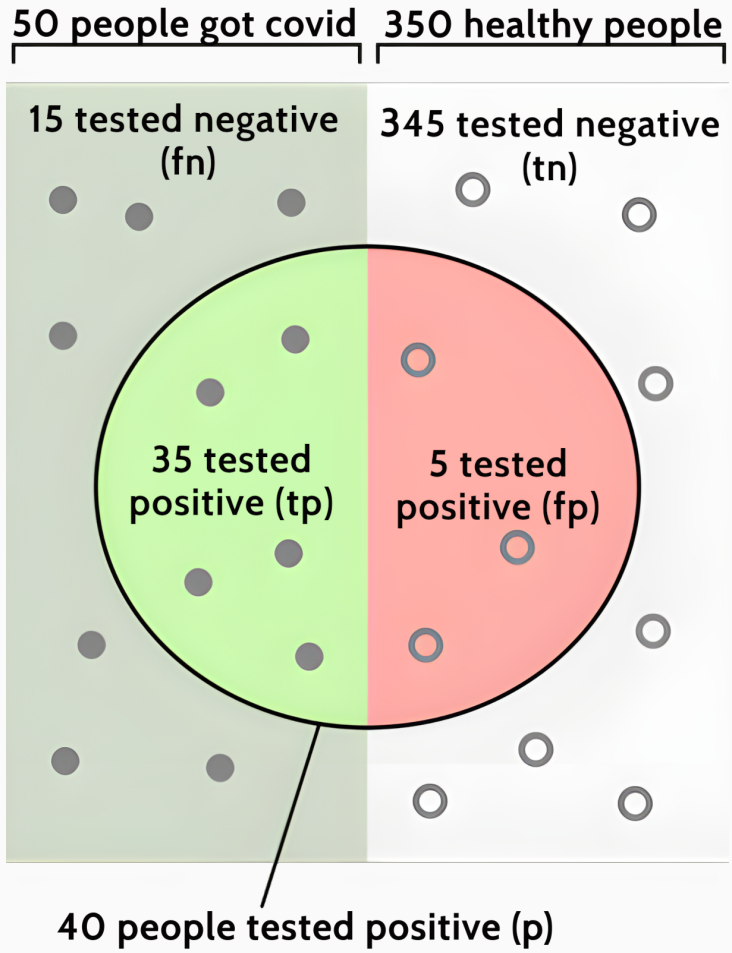
Classification framework: example

A group of 400 people did a covid-19 test. 50 people really got covid (relevant), the other 350 people do not. The test is positive for 40 people, out of which 35 really have covid.

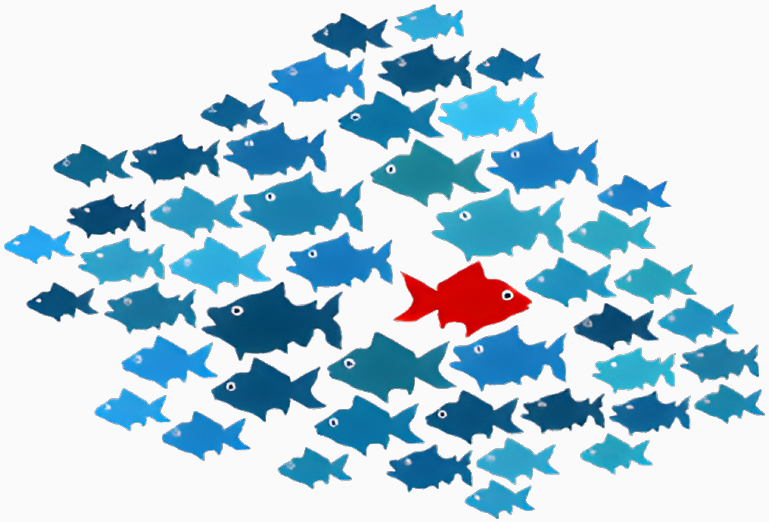
Classification framework: example

A group of 400 people did a covid-19 test. 50 people really got covid (relevant), the other 350 people do not. The test is positive for 40 people, out of which 35 really have covid.

- p: 40
- n: 360
- tp: 35
- tn: 345
- fp: 5
- fn: 15



Accuracy



$$Accuracy = \frac{tp+tn}{tp+tn+fp+fn}$$

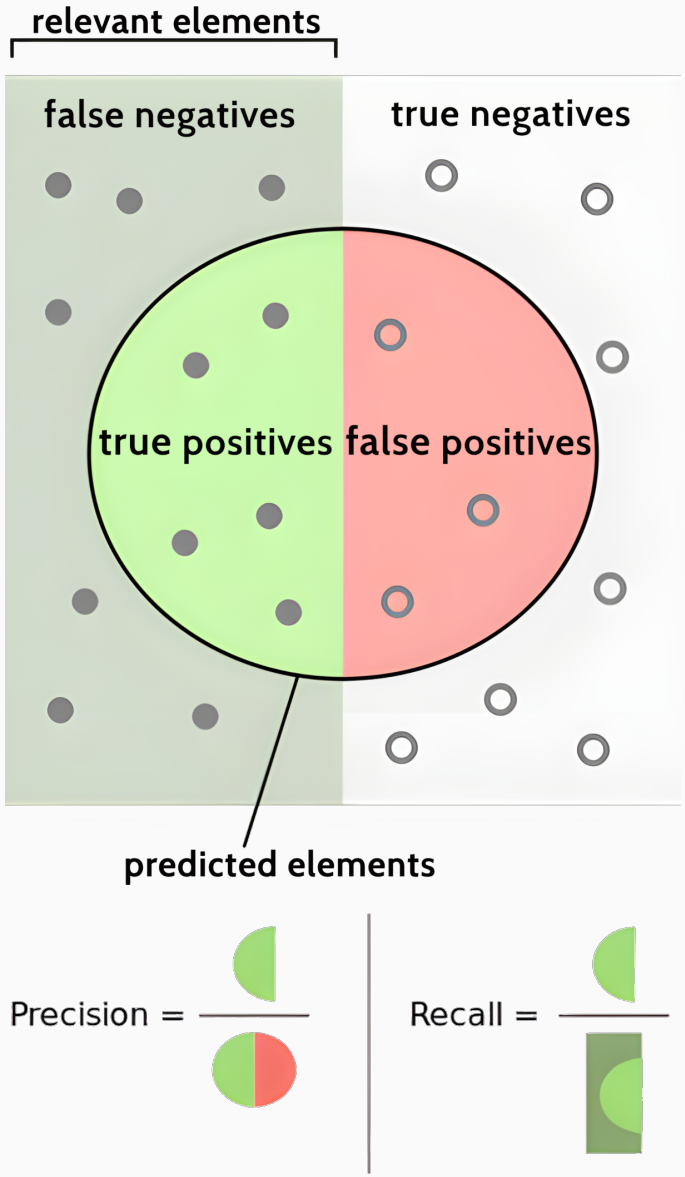
- Percentage of well classified items
- Incomplete description of the method's performance

Source: Image: Devin Soni, `towardsdatascience.com`

Precision, recall, F-score

- True positive ratios:
  - Precision
    $\rightarrow tp/(tp + fp)$
  - Recall = Sensitivity
    $\rightarrow tp/(tp + fn)$
  - Specificity:
    $\rightarrow tn/(tn + fp)$
- Complementary measures, report both
- F-score: prec. & recall harmonic mean

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$



Source: image adapted from Wikipedia

Multi-class classification

- Precision, recall and F-score are calculated for each class!
- For the case of 2 classes, often 1 is more important
  - $\rightarrow$ Spam detection: we're interested in the SPAM class
- Multi-class classification with k classes:
  - $\rightarrow$ k precision scores
  - $\rightarrow$ k recall scores
  - $\rightarrow$ k F-scores

Wooclap

Wooclap time!

Métriques de classification: exercice

Suppose you are working on a spam detection system for email. You have developed a machine learning model that has been evaluated on a test set containing 1,000 emails, 200 of which were really spam. The model classified 150 of these 200 emails as spam. 25 emails were classified as spam when they should not have been.

- Obtain the number of tp, tn, fp and fn.
- Calculate precision, recall, and accuracy.

Métriques de classification: exercice

Suppose you are working on a spam detection system for email. You have developed a machine learning model that has been evaluated on a test set containing 1,000 emails, 200 of which were really spam. The model classified 150 of these 200 emails as spam. 25 emails were classified as spam when they should not have been.

- Obtain the number of tp, tn, fp and fn.
- Calculate precision, recall, and accuracy.

- tp: 150
- tn: 775
- fp: 25
- fn: 50

- Precision: $\frac{150}{150+25} = 0.857$
- Recall: $\frac{150}{150+50} = 0.75$
- Accuracy: $\frac{150+775}{1000} = 0.925$

Accuracy and class imbalance

- Example: hate speech detection in tweets
  - Only a small percentage ($\sim$1%) are hateful
  - Let us annotate everything as not hateful
  - My model has an accuracy of 99%! So powerful!

Accuracy and class imbalance

- Example: hate speech detection in tweets
  - Only a small percentage ($\sim$1%) are hateful
  - Let us annotate everything as not hateful
  - My model has an accuracy of 99%! So powerful!

High class imbalance: accuracy is misleading/not relevant

F-score or F-measure

- F-score (or F-measure): harmonic mean of precision and recall

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

- F-score can be weighted to favour precision or recall
  - $\rightarrow$ β=0.5: More weight on precision, less weight on recall
  - $\rightarrow$ β=1: Balance the weight on precision and recall
  - $\rightarrow$ β=2: Less weight on precision, more weight on recall

$$F_\beta = (1 + \beta^2) \times \frac{Precision \times Recall}{\beta^2 Precision + Recall}$$

Other metrics (see backup slides)

- ROC curve / Area under the curve
  $\rightarrow$ Real prediction, threshold
- (Mean) average precision
  $\rightarrow$ Real prediction, binary gold classes
- Structured prediction
  $\rightarrow$ Compare trees, graphs, clusters...
- ...

Warning!

Metric bias

- Choosing the most interesting metrics after the experiment
  - $\longrightarrow$ E.g. a few positive examples, thousands of negative examples
  - $\longrightarrow$ Report accuracy, omit f-score: strongly affected by class imbalance
- Evaluation metrics must be defined before the experiment

"Just because statistical software [. . .] generates output by default does not mean that all output is useful or relevant."

Source: Bruce et al (2020)

Goodhart's law

"When a measure becomes a target, it ceases to be a good measure"

- Cobra effect
- Reinforcement learning policy
- Grade-oriented education system

- Risk: optimise evaluation metric at any expense
  - $\longrightarrow$ Overfitting, low generalisation
  - $\longrightarrow$ Forgetting the research question
  - $\longrightarrow$ Frustration with unrealistic goals
  - $\longrightarrow$ ...

Source: Thanks to François Hamonic

Evaluation metrics and statistics

- Evaluation metrics are calculated on datasets
  - $\longrightarrow$ Datasets = samples
  - $\longrightarrow$ Metrics = random variables

Evaluation metrics and statistics

- Evaluation metrics are calculated on datasets
  - $\rightarrow$ Datasets = samples
  - $\rightarrow$ Metrics = random variables

Questions

1. How can we compare these random variables?
2. Can we trust that sample-based estimations generalise?

Outline

Year 3000...

The Earth is finally a safe and pleasant place for humans again.

However, 1000 years of global warming unleash a dangerous permafrost bacteria.

The bacteria starts to infect human hosts, causing a mysterious disease.

Centuries in insipid watery ice made the bacteria obsessive about...

...vanilla ice-cream! ♡

The illness is called

- Compulsive
- Obsessive
- Vanilla
- Ice-cream
- Disease

CHAOS !!

The bacteria spreads rapidly, and infected humans start eating tons of vanilla ice-cream.

Milk prices rise to the stratosphere, ice-cream makers strike, diabetes and obesity break records...

Governments impose ice-cream lockdowns, interplanetary travel is forbidden, panic everywhere!

After months of an unprecedented crisis...

A lab finally announces a vaccine at phase 3 of evaluation!

This requires an experiment called randomized control trial

Randomized control trial

Group A Vaccine

Group B Placebo

Conclusion:

The vaccine works!

What a relief for humanity!

Average nb. ice-creams/day (ICD):

- Group A: $ICD_A = 1.47$
- Group B: $ICD_B = 1.56$

$ICD_A(\text{vaccine}) < ICD_B(\text{placebo})$

But. . . maybe humans forgot all about statistics?

- Is the observed difference large enough?
  - $ICD_A = 1.47$ ice/creams per day
  - $ICD_B = 1.56$ ice/creams per day

$$\delta = ICD_B - ICD_A = 0.09$$

- Maybe this result is due to randomness in sampling

  $\rightarrow$ Affects our conclusion that vaccine (A) better than placebo (B)?

But. . . maybe humans forgot all about statistics?

- Is the observed difference large enough?
  - $ICD_A = 1.47$ ice/creams per day
  - $ICD_B = 1.56$ ice/creams per day

$$\delta = ICD_B - ICD_A = 0.09$$

- Maybe this result is due to randomness in sampling

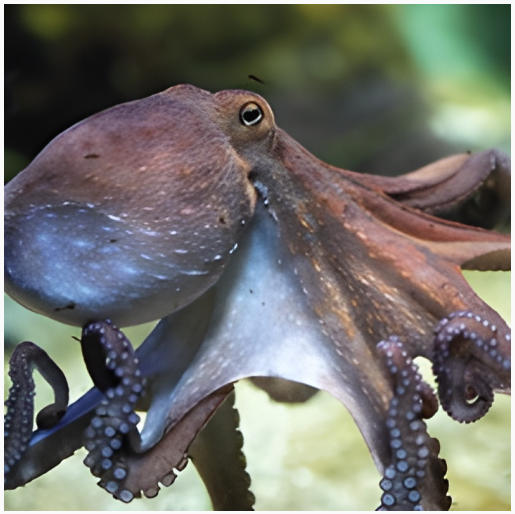
  $\rightarrow$ Affects our conclusion that vaccine (A) better than placebo (B)?

Given the samples, the metrics, and the experiment's conditions:

What is the probability of making a false claim if we conclude that A better than B in general?

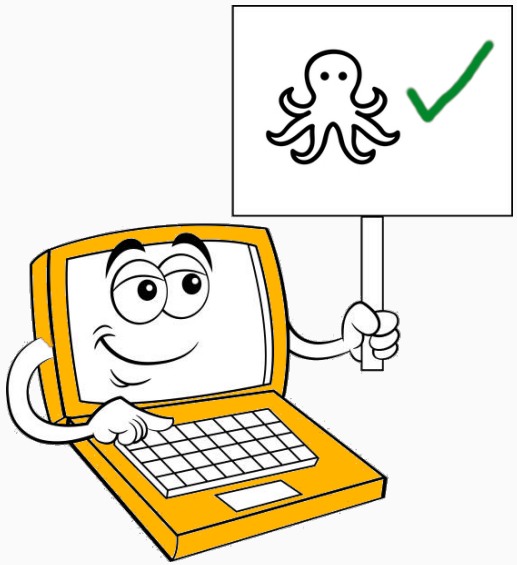
System comparison: example

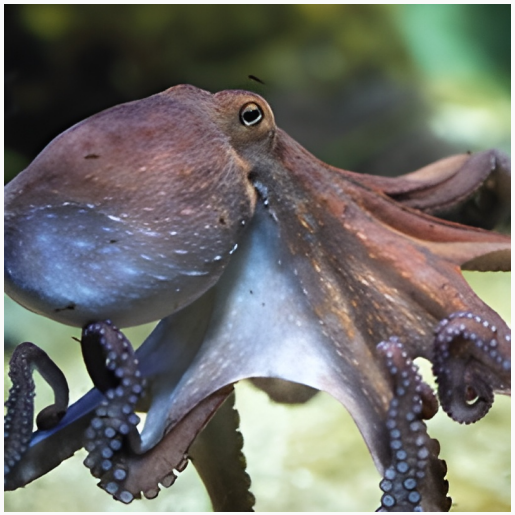
- Our Baseline system classifies images
  - $\rightarrow$ Two categories: octopus or not octopus



System comparison: example

- Our Baseline system classifies images
  - $\rightarrow$ Two categories: octopus or not octopus



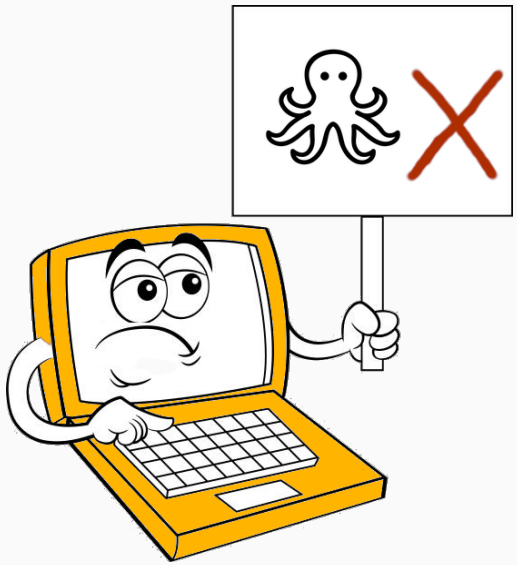


System comparison: example

- Our Baseline system classifies images
  - $\rightarrow$ Two categories: octopus or not octopus

System comparison: example

- Our Baseline system classifies images
  - → Two categories: octopus or not octopus

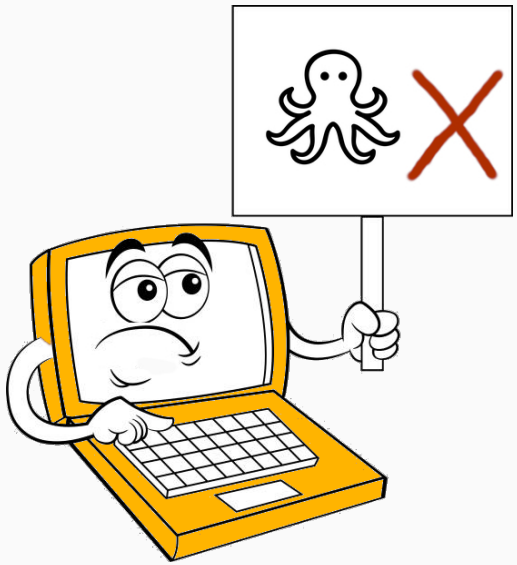


System comparison: example

- Our Baseline system classifies images
  - $\rightarrow$ Two categories: octopus or not octopus

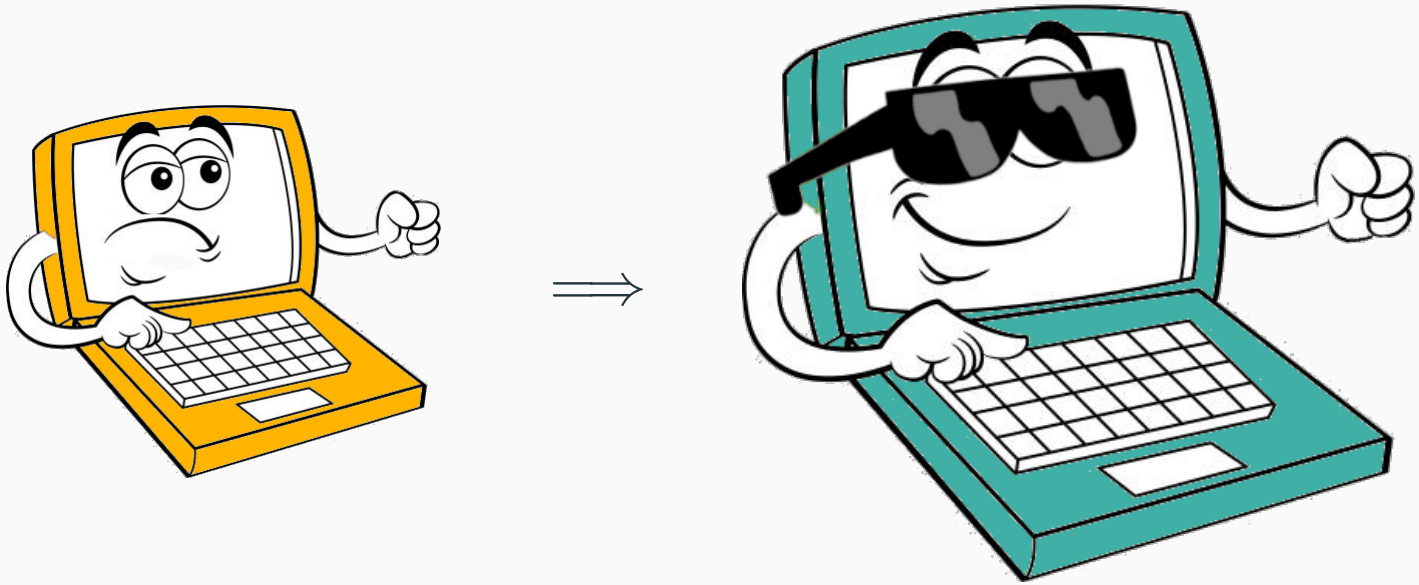
System comparison: example

- Our Baseline system classifies images
  - → Two categories: octopus or not octopus
- Sometimes it makes mistakes



System comparison: example

- We developed an Awesome new system!
  - $\rightarrow$ E.g. the new system was trained on more data

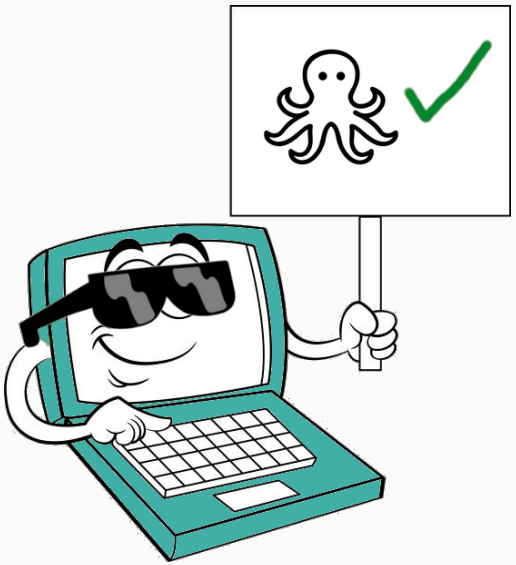


System comparison: example

- We developed an Awesome new system!
  - $\rightarrow$ E.g. the new system was trained on more data

System comparison: example

- We developed an Awesome new system!
  - $\rightarrow$ E.g. the new system was trained on more data
- It seems that it makes less mistakes $\Longrightarrow$ 🎉

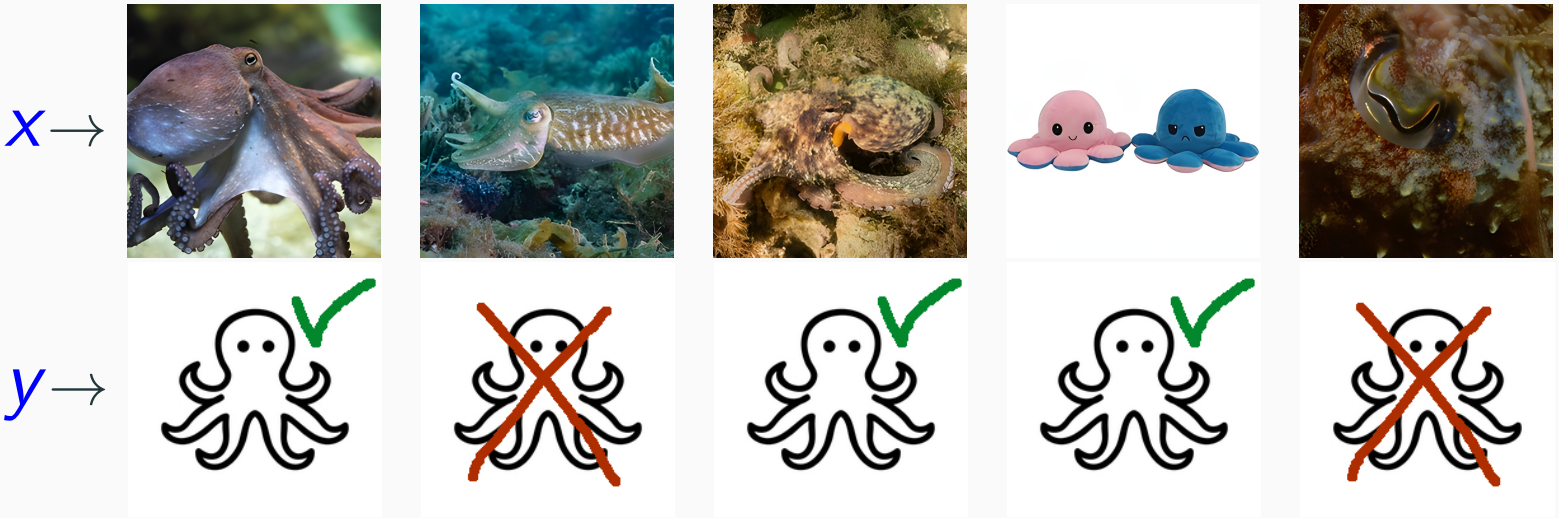


Test set

- Is A really better than B?
  - → Testing on a few examples is not enough!
- Use a test set containing (x,y) pairs
  - → x - sea animal images
  - → y - reference octopus/other labels
- The test set was not used to develop the system

Test set: example

Images x selected to be in the held-out test set

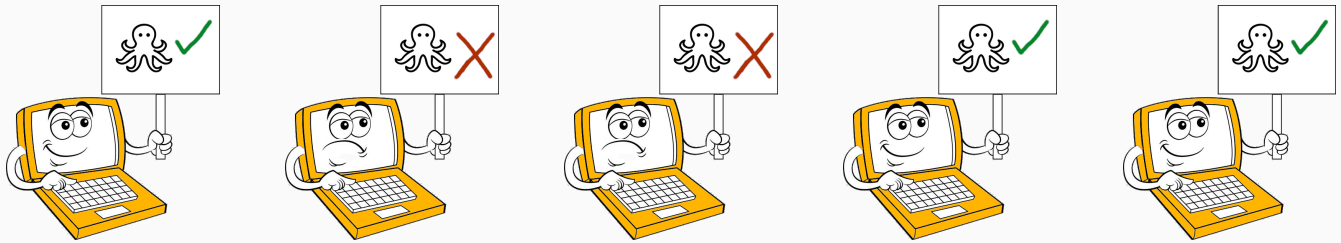


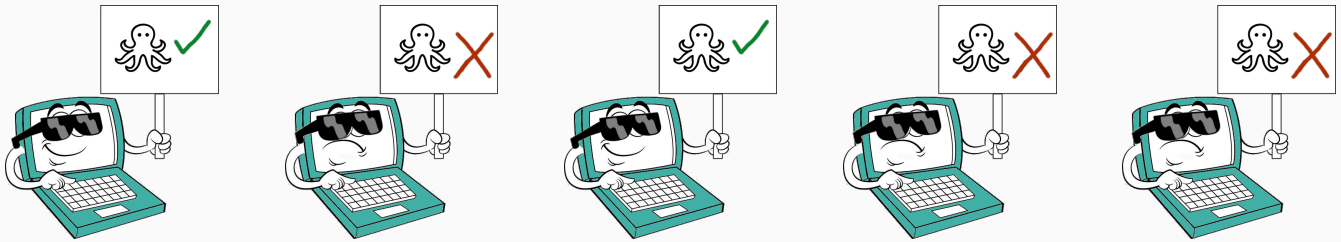
Gold labels y considered as reference (e.g. annotated by humans)

System predictions

Both systems generate predictions $\hat{y}$ for test set instances x

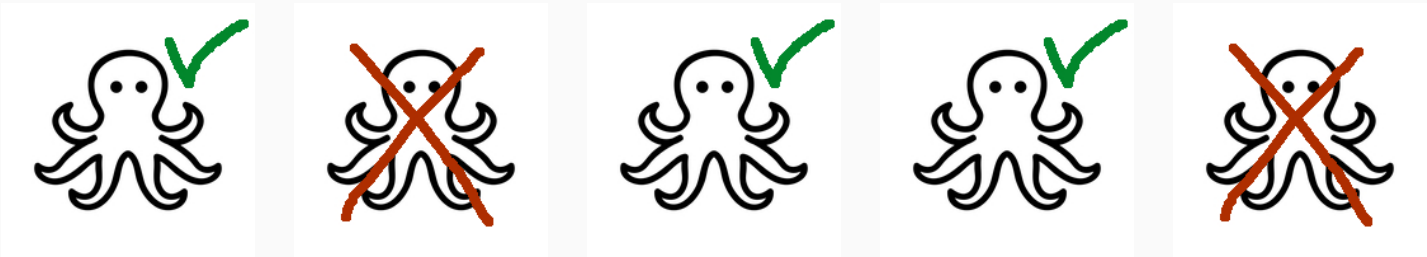
$x \rightarrow$ 

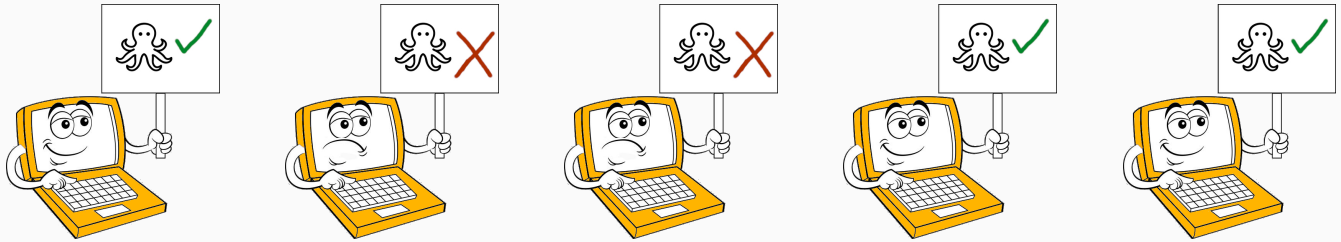
$\hat{y}_B(x) \rightarrow$ 

$\hat{y}_A(x) \rightarrow$ 

Evaluation metrics

Compare predictions $\hat{y}_B$ and $\hat{y}_A$ to reference y

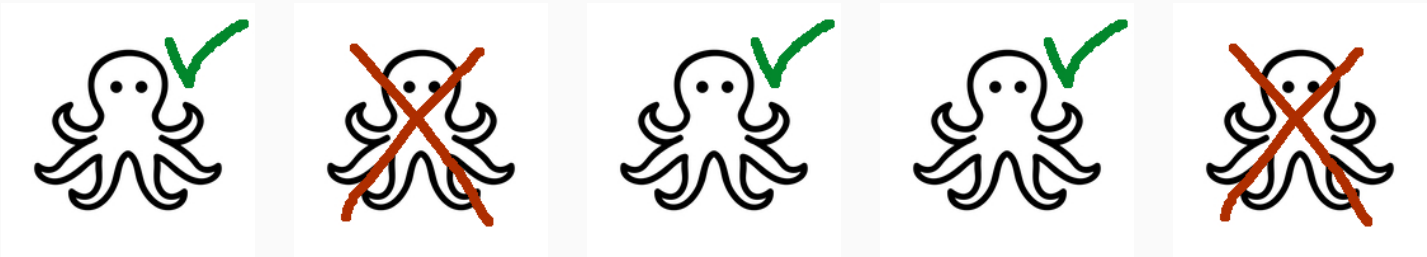
$y \rightarrow$ 

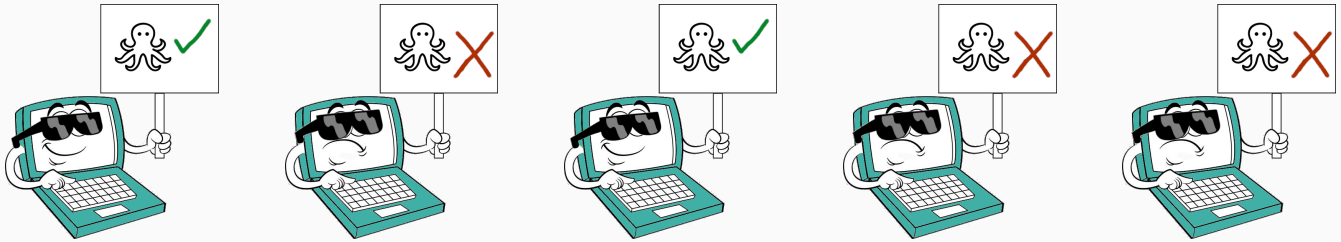
$\hat{y}_B(x) \rightarrow$ 

$$M(B, x, y) = \frac{3}{5} = 0.6$$

Evaluation metrics

Compare predictions $\hat{y}_B$ and $\hat{y}_A$ to reference y

$y \rightarrow$ 

$\hat{y}_A(x) \rightarrow$ 

$$M(A, x, y) = \frac{4}{5} = 0.8$$

Wooclap

Wooclap time!

System score comparison

- The accuracies of both systems are:

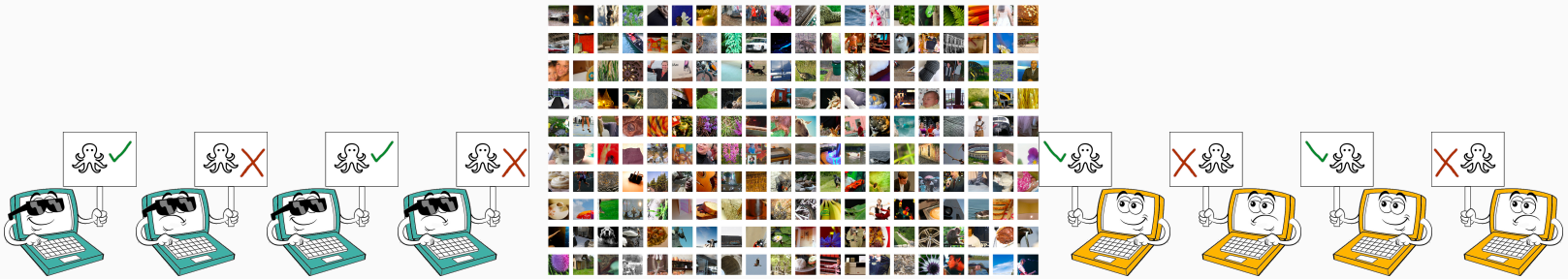
$$M(B, x, y) = \frac{3}{5} = 0.6$$
$$M(A, x, y) = \frac{4}{5} = 0.8$$

- It seems like A is better than B
- The difference (delta) is positive

$$\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y) = 0.8 - 0.6 = \boxed{0.2}$$

System comparison: example

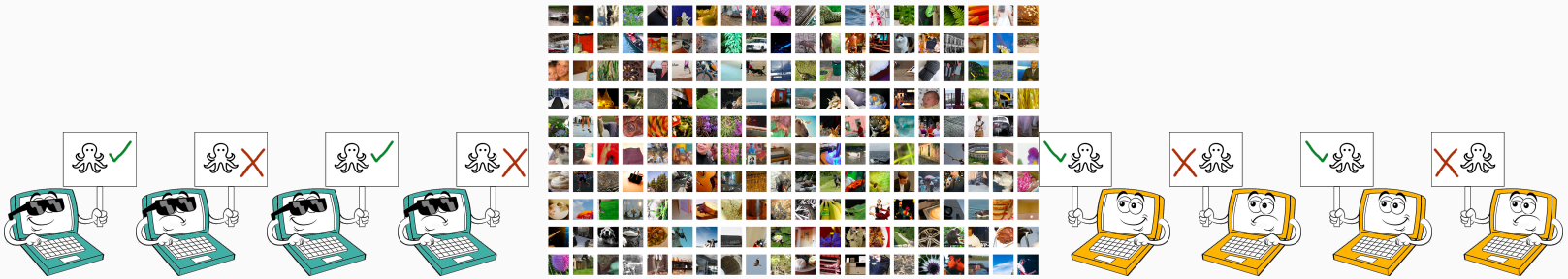
We obtained a much larger test set x',y'



We compare A and B again and obtain:

System comparison: example

We obtained a much larger test set x',y'

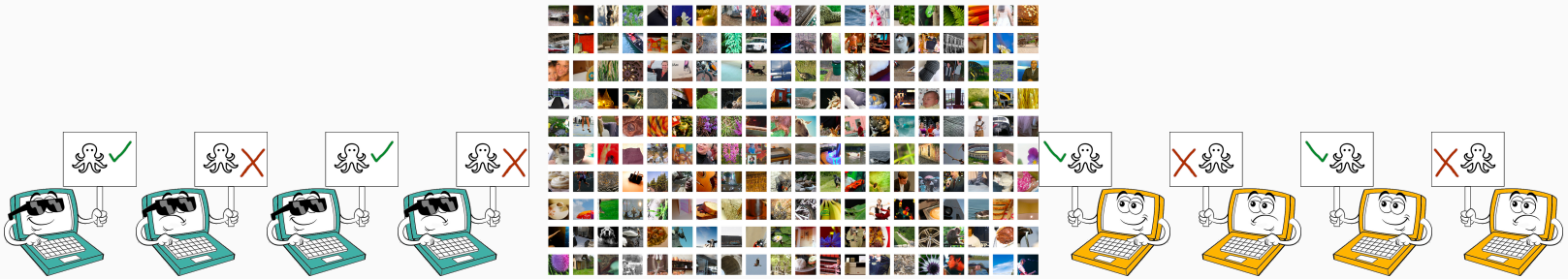


We compare A and B again and obtain:

$$\begin{aligned}\delta_{A-B}(x', y') &= M(A, x', y') - M(B, x', y') \\ &= 0.7612 - 0.7586 \\ &= \boxed{0.0026}\end{aligned}$$

System comparison: example

We obtained a much larger test set x',y'



We compare A and B again and obtain:

$$\begin{aligned}\delta_{A-B}(x', y') &= M(A, x', y') - M(B, x', y') \\ &= 0.7612 - 0.7586 \\ &= \boxed{0.0026}\end{aligned}$$

- Can we still affirm that A is better than B?
- If we add or remove a couple of images, could the result flip?

Interpreting delta

$$\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y)$$

- Delta: compare both systems with a single value
  - $\rightarrow$ A better than B $\rightarrow$ $\delta_{A-B}(x, y) > 0$
  - $\rightarrow$ A equivalent to B $\rightarrow$ $\delta_{A-B}(x, y) = 0$
  - $\rightarrow$ A worse[3] than B $\rightarrow$ $\delta_{A-B}(x, y) < 0$
- In some disciplines, $\delta_{A-B}(x, y)$ is called effect

[3] Yes, the old Baseline may outperform the new Awesome system!

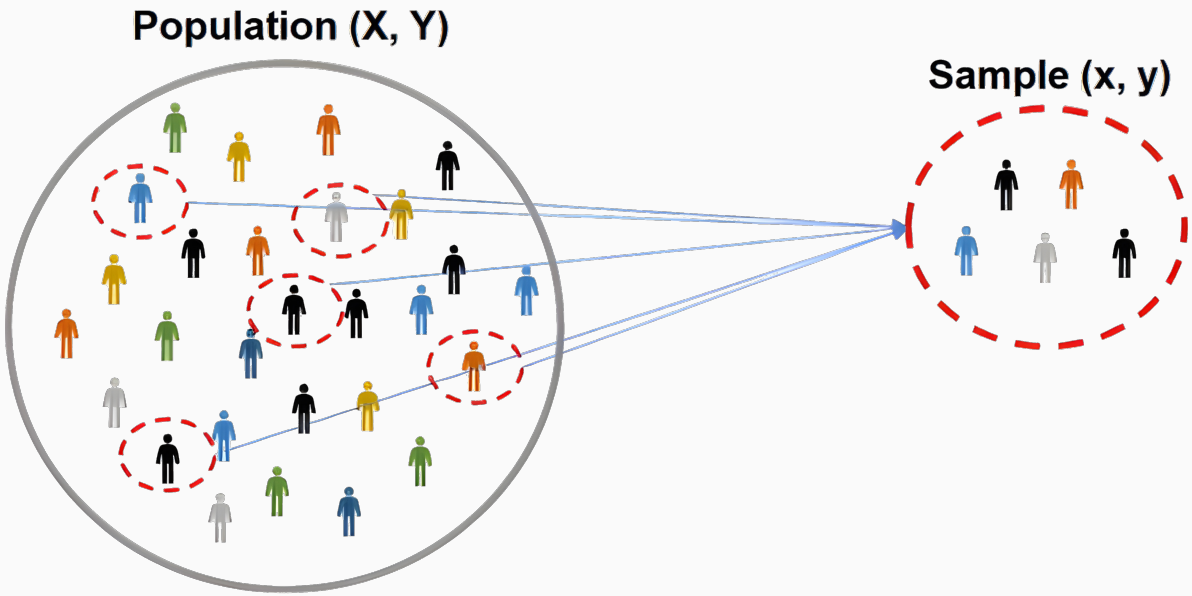
System comparison in a nutshell

1. We develop a system A supposed to be better than B
2. To verify this, we apply both systems to the same test set:
   - $\rightarrow$ Get output of system A on the test set (x, y)
   - $\rightarrow$ Get output of system B on the test set (x, y)
3. Calculate the evaluation metric $M(\cdot)$ for both outputs

$$\delta_{A-B}(x, y) = M(A, x, y) - M(B, x, y)$$

4. Large positive $\delta_{A-B}(x, y) \implies$ 🎉
5. In practice, $\delta_{A-B}(x, y)$ is often small 😎

Test sets as random samples

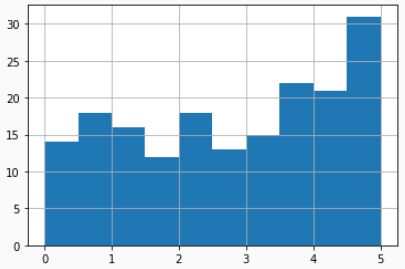


- Could the observed $\delta_{A-B}(x, y) > 0$ be due to sampling error ?
  - $\rightarrow$ (x, y) is a sample of joint random variables (X, Y)
  - $\rightarrow$ What difference/effect would we get for another sample (x', y') ?
- What is the probability that A is actually no better than B
  - $\rightarrow$ If we ever had access to the "real" distribution of (X, Y)?

Effects as random variables

- We obtain a single $\delta_{A-B}(x, y)$ value
- This value depends on the test set (x, y), which is a sample
- We see $\delta_{A-B}(x, y)$ as a sampled value of a random variable

$$\delta_{A-B}(X, Y) \rightsquigarrow$$



P-value

- **P-value:** probability of obtaining at least $\delta_{A-B}(x, y)$
  - When in reality, A is no better than B

P-value

- **P-value:** probability of obtaining at least $\delta_{A-B}(x, y)$
  - When in reality, A is no better than B

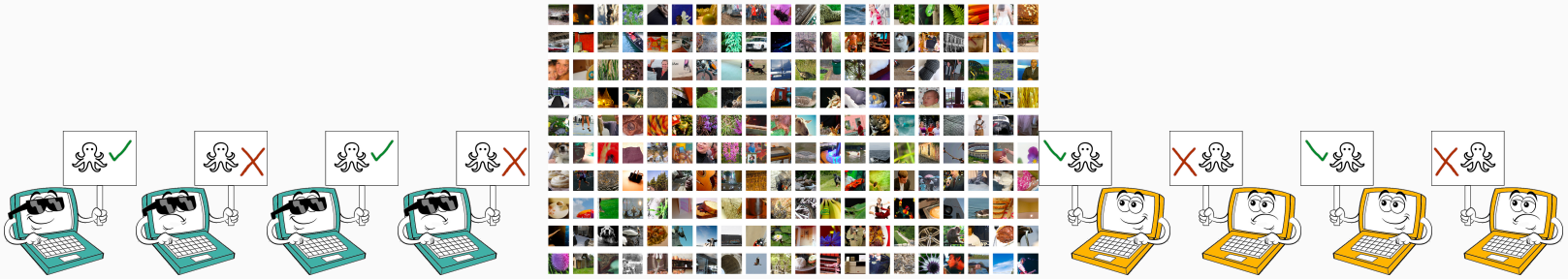
Warning!

- p-value $\neq$ probability that your conclusion is wrong!
- Rather, probability that chance is the only reason for the delta

Wooclap

Wooclap time!

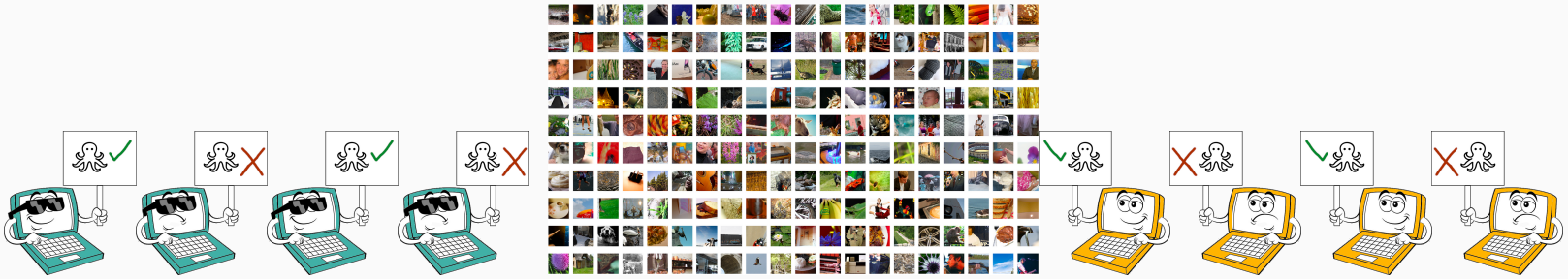
P-value: example



We have one delta value obtained on the large dataset (x', y')

$$\delta_{A-B}(x', y') = 0.0026$$

P-value: example



We have one delta value obtained on the large dataset (x', y')

$$\delta_{A-B}(x', y') = 0.0026$$

If we had all possible images of sea creatures and their classes

$\rightarrow$ Imagine we have access to the real distribution $\delta_{A-B}(X, Y)$

- P-value = probability of obtaining 0.0026 difference (or more)
- If A is actually no better than B

Hypothesis testing

- H_0: $\delta_{A-B}(X, Y) \leq 0 \implies$ if true, then A not better than B
- H_1: $\delta_{A-B}(X, Y) > 0$

- Goal: reject H_0

  $\rightarrow$ Conclusion: significant difference between the systems

Hypothesis testing and p-value

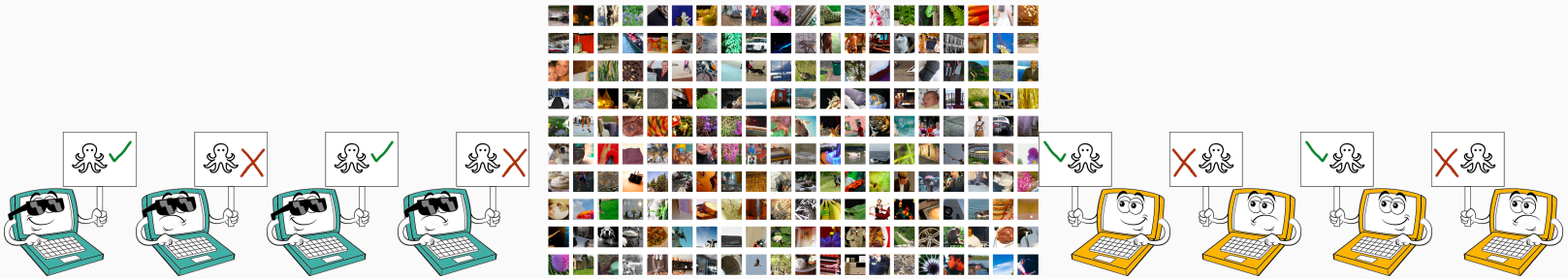
Hypotheses

- H_0: $\delta_{A-B}(X, Y) \leq 0$
- H_1: $\delta_{A-B}(X, Y) > 0$

- **P-value:** probability of observing $\delta_{A-B}(x, y)$ while H_0 true

  $\longrightarrow$ Intuition: if H_0 is true, large $\delta_{A-B}(x, y)$ is unlikely

Hypothesis testing: example



$$\text{p-value} = P\{\delta_{A-B}(X, Y) \geq 0.0026\} \text{ assuming } \delta_{A-B}(X, Y) \leq 0$$

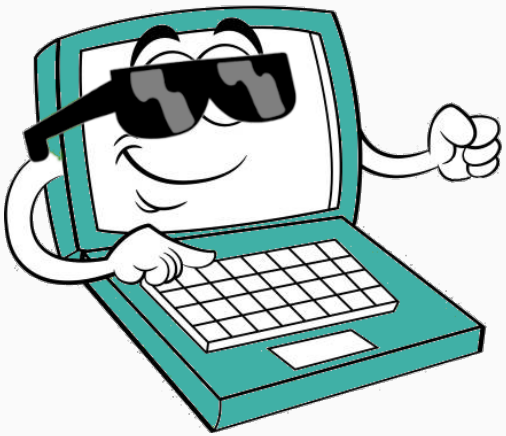
Usually, if p-value small enough $\implies$ we conclude A better than B

Type I errors

- Type I error: false positive

  $\rightarrow$ Rejecting H_0 when it is actually true

Conclusion of the test:



is better than

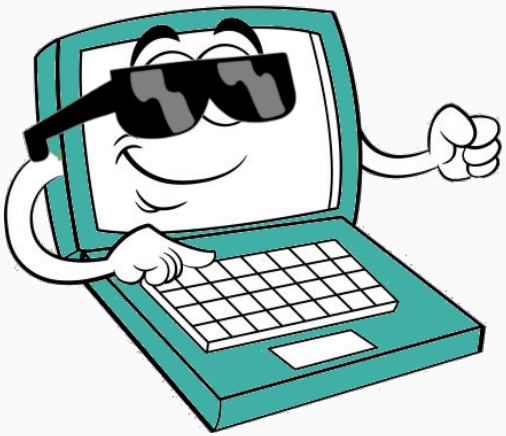
Reality: But it isn't better !

Type II errors

- Type II error: false negative

  $\rightarrow$ Not rejecting H_0 when it is actually false

Conclusion of the test:



is not better than

Reality: But it is better !

Goal

- Probability of type-I error is upper bounded by α

  $\rightarrow$ α is called the significance level or threshold

- Probability of type-II error is as low as possible

  $\rightarrow$ Test power: ability to avoid type-II errors

Statistically significant result

$$\text{p-value} < \alpha \implies \text{statistically significant!}$$ 🎉

- p-value: probability of extreme outcome under null hypothesis
- α: significance level (threshold)

  $\rightarrow$ Usual "magic" value: $\alpha = 0.05$

Statistically significant result

$$\text{p-value} < \alpha \implies \text{statistically significant!}$$ 🎉

- p-value: probability of extreme outcome under null hypothesis
- α: significance level (threshold)

  $\rightarrow$ Usual "magic" value: $\alpha = 0.05$

The word significant should not be used for anything else

Wooclap

Wooclap time!

What's a "good" p-value?

- We want the p-value to be below α (significance level)
  - $\rightarrow$ Usual threshold: $\alpha = 0.05$
  - $\rightarrow$ p-value=0.045 and p-value=0 are below the α threshold
- If so, we say that the A-B difference is significant

How can we estimate p-values?

- P-value depends on $\delta_{A-B}(X, Y)$ probability distribution
- Which in turn depends on $M(A, x, y)$ and $M(B, x, y)$

  $\rightarrow$ Remember: $M(\cdot)$ is our evaluation metric
- $M(\cdot)$'s distribution determines that of δ

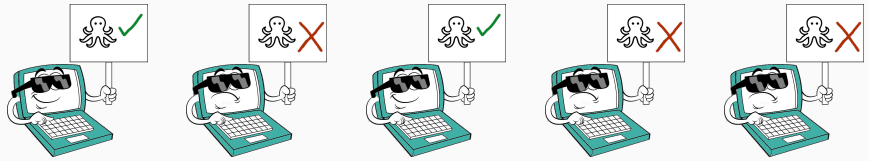
  $\rightarrow$ Study the probability distribution of evaluation metric $M(\cdot)$

Accuracy is an average

$y \rightarrow$ 

$\hat{y}_B \rightarrow$ 

1 1 0 1 0

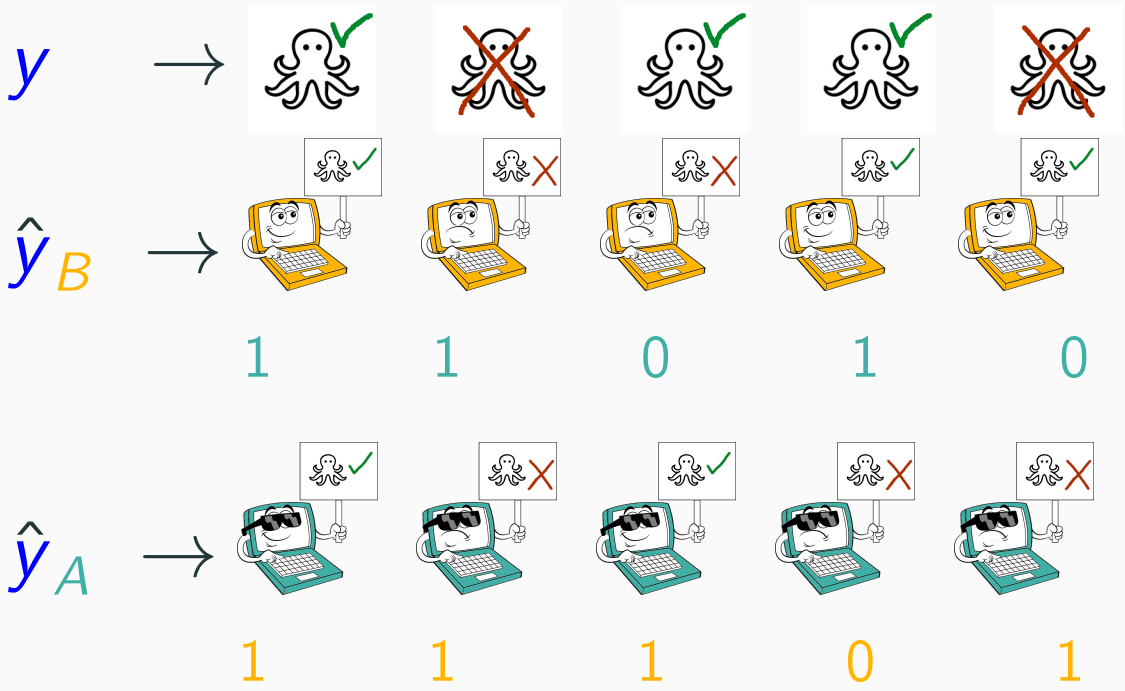
$\hat{y}_A \rightarrow$ 

1 1 1 0 1

$Acc_B = \frac{1+1+0+1+0}{5} = \frac{3}{5}$ $\quad Acc_A = \frac{1+1+1+0+1}{5} = \frac{4}{5}$

Accuracy is an average

Accuracy is an average



$$Acc_B = \frac{1+1+0+1+0}{5} = \frac{3}{5} \qquad Acc_A = \frac{1+1+1+0+1}{5} = \frac{4}{5}$$

Accuracy is an average $\implies$ normally distributed!

$\rightarrow$ Thanks to the central limit theorem

The t-test for paired samples

- T-test: hypothesis testing for normally distributed variables
  → Based on Student's t distribution (similar to normal)

$$\text{t-stat} = \frac{M(A, x, y) - M(B, x, y)}{SE/\sqrt{m}}$$

- m: size of the paired sample (x, y)
- SE: standard deviation of difference $d = [\hat{y}_A = y] - [\hat{y}_B = y]$
- P-value: look up Student's t table, $m - 1$ degrees of freedom
  → E.g. `https://homepage.divms.uiowa.edu/~mbognar/applets/t.html`
  → In practice: scipy `stats.ttest_rel`

T-test toy example (gory details)

$$M(A, x, y) = \frac{1+1+1+0+1}{5} = 0.8 \qquad M(B, x, y) = \frac{1+1+0+1+0}{5} = 0.6$$

$$d = [\hat{y}_A = y] - [\hat{y}_B = y] = (0, 0, 1, -1, 1) \qquad \overline{d} = 0.2$$

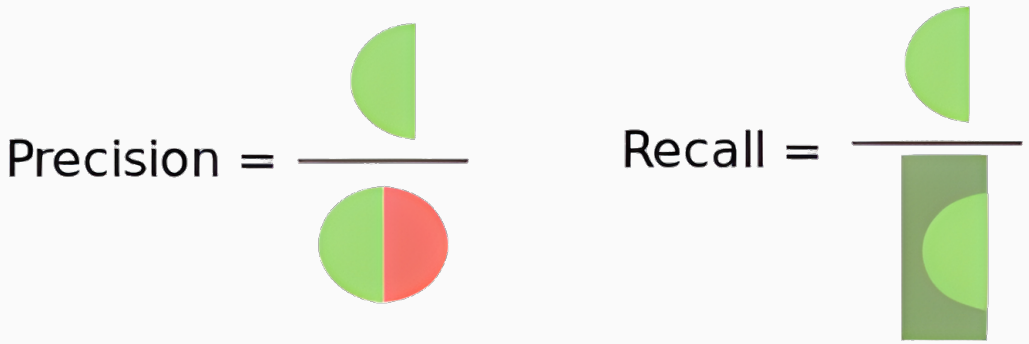
$$SE = \sqrt{\frac{(0-.2)^2+(0-.2)^2+(1-.2)^2+(-1-.2)^2+(1-.2)^2}{5-1}} = \sqrt{\frac{2.8}{4}} = 0.837$$

$$\begin{aligned} \text{t-stat} &= \frac{M(A, x, y) - M(B, x, y)}{SE/\sqrt{m}} \\ &= \frac{0.8 - 0.6}{0.837/\sqrt{5}} \\ &= \mathbf{0.5345} \end{aligned}$$

$$\boxed{\text{p-value} = P(X > 0.5345) = \mathbf{0.3107}}$$

System A not significantly better than B at $\alpha = 0.05$ level!

Non parametric tests



- Problem of t-test : assumes $M(A, x, y) \sim$ normally distributed
- Other metrics :
  - Recall $R = \frac{tp}{tp+fn}$, $tp + fn$ constant
    $\rightarrow$ t-test OK ✓
  - Precision $P = \frac{tp}{tp+fp}$ depends on $tp + fp$, unknown distribution
    $\rightarrow$ t-test not OK ✗
  - F-score $= \frac{2 \cdot P \cdot R}{P+R}$ depends on P, unknown distribution
    $\rightarrow$ t-test not OK ✗

Parametric vs. non parametric

Many authors use the terms parametric vs. non parametric tests

- What does it mean?

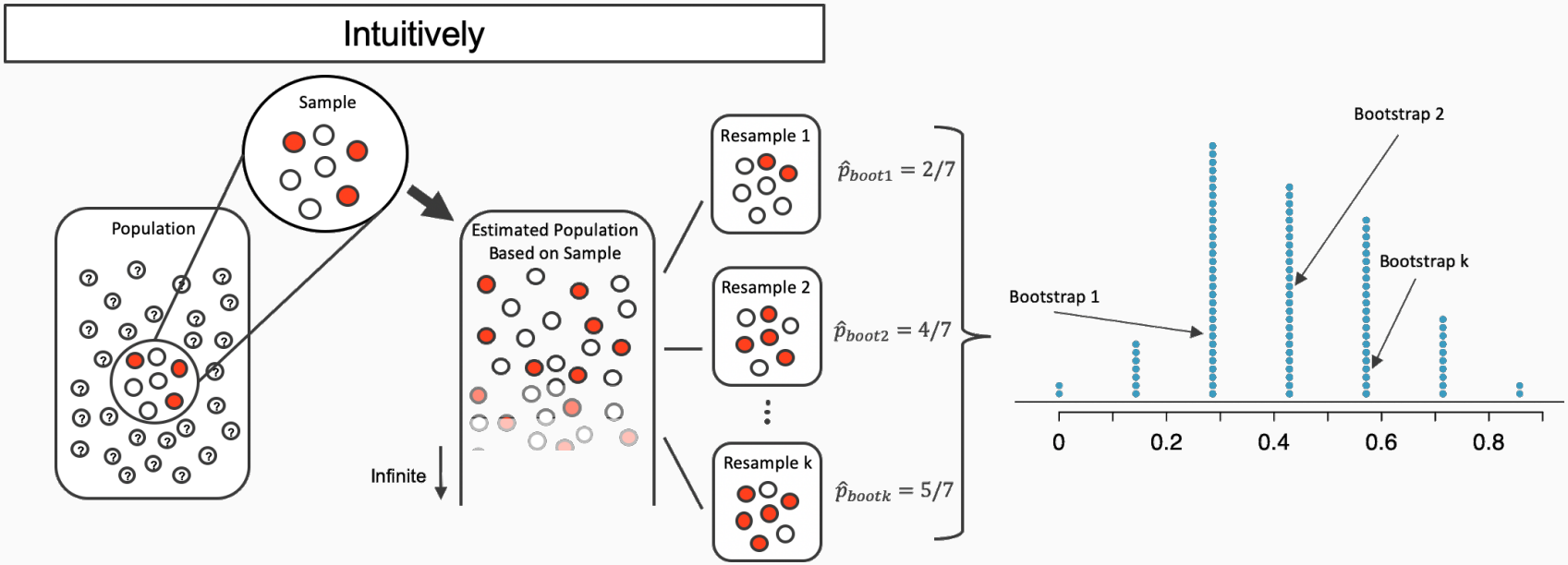
Most of the time, "parametric" means:
The underlying random variable is normally distributed

Non parametric tests

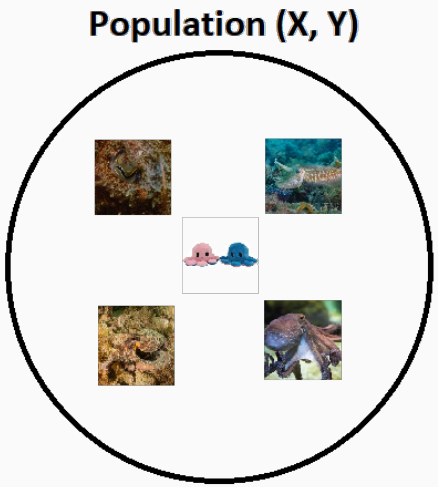
- Alternative : non parametric tests
  1. No sampling
     - Fast
     - Conservative, A no better than B for small δ (low power)
     - E.g. sign test, McNemar, Wilcoxon
  2. With sampling
     - Slow
     - Powerful, rejects H_0 more often (low type-II error probability)
     - E.g. permutation (randomised approximation), bootstrap

Bootstrap

Idea : estimate M distribution by random re-sampling in (x, y)



Poolpstrap



$$Acc(A, \text{Sample 1}) = \frac{2}{2}$$
$$Acc(B, \text{Sample 1}) = \frac{1}{2}$$

$$Acc(A, \text{Sample 2}) = \frac{1}{2}$$
$$Acc(B, \text{Sample 2}) = \frac{1}{2}$$

.
.
.

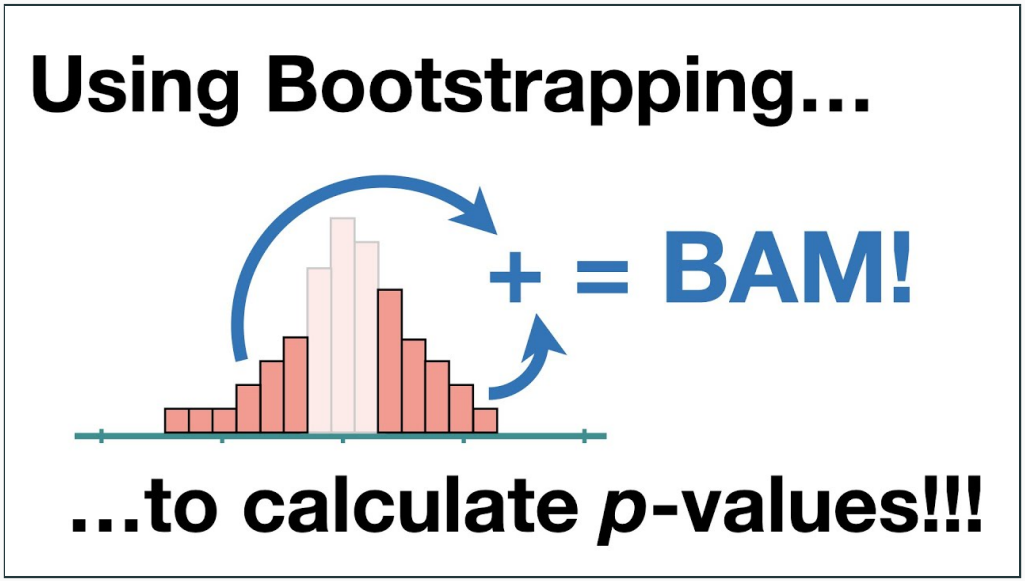
$$Acc(A, \text{Sample k}) = \frac{2}{2}$$
$$Acc(B, \text{Sample k}) = \frac{2}{2}$$

Bootstrap for significance

```
1  deltaobs = M(A,x,y) - M(B,x,y)   # delta on test set
2  R = 10000                        # 10k random samples
3  for i = 1 .. R :
4    xs, ys = sample(x,y,m)         # with repetition
5    deltasample = M(A,xs,ys) - M(B,xs,ys)
6    if deltasample > 2 * deltaobs :
7          r = r + 1
8  pvalue = r/R
```

Why comparing with 2 $\times$ deltaobs?

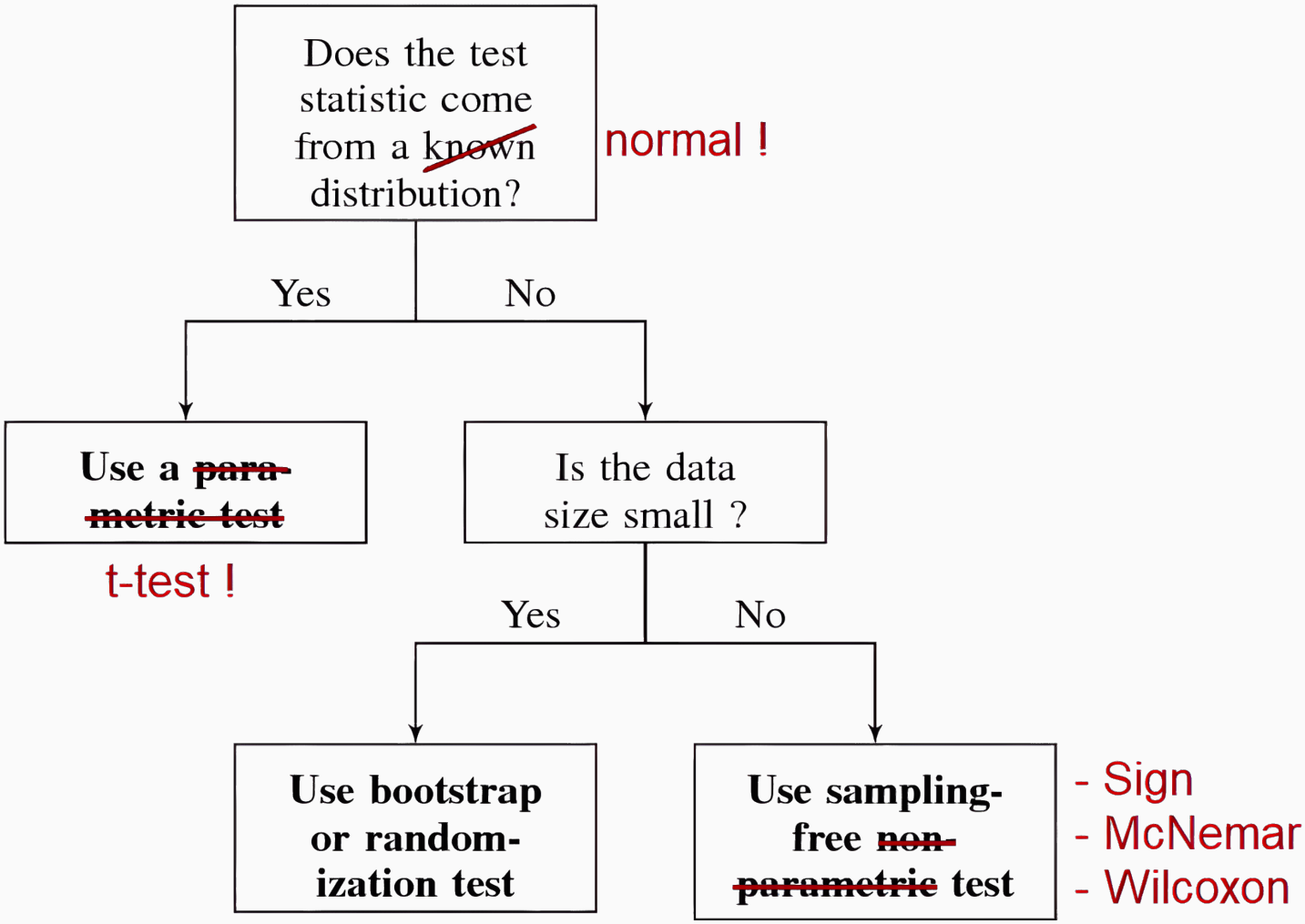
- Overall sample mean is $\delta_{A-B}(x, y)$
  - $\rightarrow$ Standardize bootstrat sample (x', y') to assume H_0 (zero effect)
  - $\rightarrow$ $H_0 \implies$ substract observed delta from each bootstrap sample value
  - $\rightarrow$ $\delta_{A-B}(x', y')$ - $\delta_{A-B}(x, y) > \delta_{A-B}(x, y)$
- Details: `https://www.youtube.com/watch?v=N4ZQQqyIf6k`



Bootstrap and hypothesis testing

- Parametric tests were more popular in the past
  - $\rightarrow$ No computer power for intensive resampling
- Today, non-parametric tests like bootstrap are most popular
  - $\rightarrow$ No need to assume normality
- Bootstrap: easy system A and B comparison
  - $\rightarrow$ Same test set $(x, y) \implies$ paired sample

Which test to apply?



Source: adapted from Dror et al. (2018)

Community's practice

Computational linguistics conferences (ACL) and journals (TACL)

General Statistics	**ACL '17**	**TACL '17**
Total number of papers	196	37
# papers that **do not** report significance	117	15
# papers that report significance	63	18
# papers that report significance but use the **wrong** statistical test	6	0
# papers that report significance but do not mention the test name	21	3

Source: Dror et al. (2018)

Outline

Error analysis

- Characterise the errors in our system's output
- Scripts to print characteristics of errors
  - $\rightarrow$ Frequency, length, resolution, predicted/gold class, ...
  - $\rightarrow$ Example: elements whose prediction is furthest from gold value
- Manual error annotation: taxonomies, guidelines
  - $\rightarrow$ Gain insight on most promising improvements

Interpretability analysis

Try to understand why systems generate a prediction

- Feature-based methods (SHAP, LIME)
  → Which parts of the inputs influence prediction?
- Visualisation
  → Attention salience, 2-D projection (UMAP, t-SNE, topology)
- Adversarial examples, perturbations
  → Difficult minimal pairs

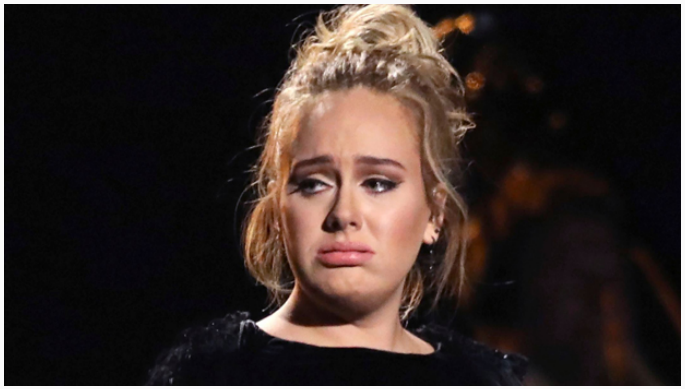
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Source: `https://homes.cs.washington.edu/~marcotcr/blog/lime/`

Negative results

- Well designed hypotheses $\rightarrow$ interesting "negative" results
- Experiments require persistence and some faith
- Source of frustration: publish or perish

  $\rightarrow$ Is it a problem with my results or with the system?
- Negative results are publishable if sound experimental design



Leaderboards

- Remember Goodhart's law (metric $\neq$ objective)
- Beating state of the art is good
- Learning something interesting about the problem is better
- From time to time: remember the research question

Thanks!

That's all for today

Carlos Ramisch
first.last@lis-lab.fr

M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

Sources i

- Adeline Paiement's course *Initiation à la recherche*
- Bruce et al. (2020) *Practical Statistics for Data Scientists*, 2nd Ed.
- Lanhenke (2022) *Understanding Random Variables and Probability Distributions*
- Yibi Huang's statistics course `https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf`
- Wikipedia *Precision and recall* and *Spearman Correlation*
- Youtube channels: *DATAtab*, *StatQuest*
- Feedback from participants of previous course editions
- Slides illustrated with the help of: Google images, `imgupscaler.com`, Canva

Sources ii

- Slides written with the help of: ChatGPT, Google Bard, DeepL, Linguee, Overleaf
- Funding: French ANR, through SELEXINI project (ANR-21-CE23-0033-01)

Backup slides

Random variables: formal definition i

- Experiment: flip 3 different coins, note head (H) or tail (T)
- The sample space S contains all possible experiment outcomes
  - $\rightarrow$ The subsets of S are called events E_i
- The random variable X denotes the number of heads (H)
  - A variable whose exact value is unknown or irrelevant
  - We know (or estimate) its probability distribution $P\{X = x_i\}$

E_i	$\{HHH\}$	$\{THH, HTH, HHT\}$	$\{TTH, THT, HTT\}$	$\{TTT\}$
$P(E_i)$	$1/8$	$1/8 + 1/8 + 1/8$	$1/8 + 1/8 + 1/8$	$1/8$
X	0	1	2	3
$P\{X = x_i\}$	$1/8$	$3/8$	$3/8$	$1/8$

Random variables: formal definition ii

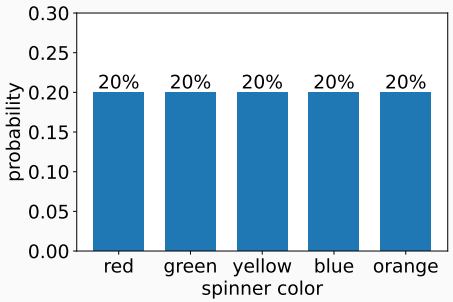
Formalisation

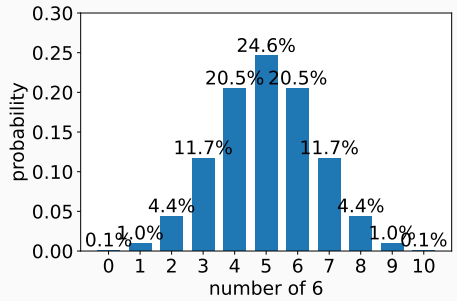
A random variable is a function $X : S \to \mathbb{R}$ such that :

1. Discrete random variable:
   - $\to$ Its set of possible values $X(S) = \{x_i, i \in \mathbb{N}^*\}$ is countable
   - $\to$ For all $x_i \in X(S)$: $\{X = x_i\} \Leftrightarrow \{e_i \in S | X(e_i) = x_i\} \in \mathcal{F}$
   - $\to$ $\mathcal{F}$ is the set of all possible events (subsets) of S
   - $\to$ $p(x_i) = P\{X = x_i\}$ is the probability mass function of X
2. Continuous random variable:
   - $\to$ $\forall$ value $x \in (-\infty, +\infty)$, $\forall$ interval $B \in \mathbb{R}$
   - $\to$ A non-negative function $P\{X \in B\} = \int_B f(x)\,dx$ exists
   - $\to$ $f(x)$ is the probability density function of X

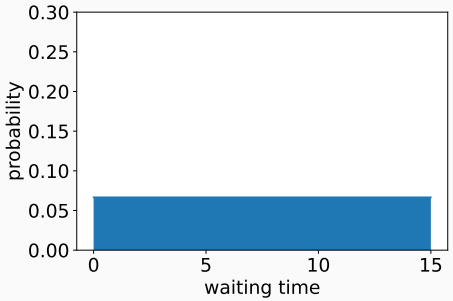
Types of probability distributions

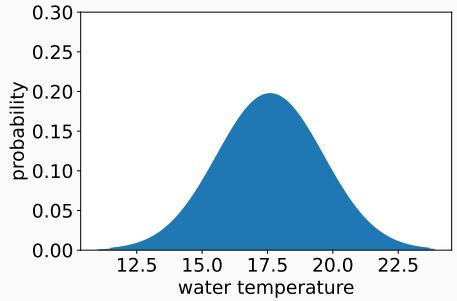
- Discrete random variables
  - → Bar graphic, finite set of values
  - → Probability at exact value $P\{X = a\}$





- Continuous random variables
  - → Line graphic, uncountable set of values (real numbers)
  - → Probability of interval $P\{a < X < b\}$





Random sample or i.i.d. variables?

- Sampled items can be seen as n random variables $X_1 \dots X_n$
  - $\rightarrow$ For instance, tossing a coin n times
- We assume that all variables have the same distribution
- We assume that all items are independent[4]
- This is often stated as independent and identically distributed
  - $\rightarrow$ The acronym i.i.d. is usually employed in probability

[4] Formally: $\forall X_i \neq X_j, \forall a, b \in X_i(S) \;\; P\{X_i = a | X_j = b\} = P\{X_i = a\}$

Random sample or i.i.d. variables?

- Sampled items can be seen as n random variables $X_1 \dots X_n$
  - $\rightarrow$ For instance, tossing a coin n times
- We assume that all variables have the same distribution
- We assume that all items are independent[4]
- This is often stated as independent and identically distributed
  - $\rightarrow$ The acronym i.i.d. is usually employed in probability

Random sample = set of n values of i.i.d. variables $X_1 \dots X_n$

[4] Formally: $\forall X_i \neq X_j, \forall a, b \in X_i(S) \;\; P\{X_i = a | X_j = b\} = P\{X_i = a\}$

Getting to the variance

Idea 1: average the difference between each value and the mean

$$\sum_{i=1}^{n} \frac{x_i - \overline{x}}{n}$$

- Calculate this amount for the sample `[-4, -4, 4, 4]`

Getting to the variance

Idea 1: average the difference between each value and the mean

$$\sum_{i=1}^{n} \frac{x_i - \overline{x}}{n}$$

- Calculate this amount for the sample `[-4, -4, 4, 4]`

$$\frac{(-4-0)+(-4-0)+(4-0)+(4-0)}{4} = 0 \quad \text{☹}$$

Getting to the variance

Idea 2: average the absolute value of the $x_i - \overline{x}$ difference

$$\sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$$

- Calculate this amount for the sample `[-4, -4, 4, 4]`

Getting to the variance

Idea 2: average the absolute value of the $x_i - \overline{x}$ difference

$$\sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$$

- Calculate this amount for the sample `[-4, -4, 4, 4]`

$$\frac{|-4-0| + |-4-0| + |4-0| + |4-0|}{4} = 4 \quad \text{☺}$$

Getting to the variance

Idea 2: average the absolute value of the $x_i - \overline{x}$ difference

$$\sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$$

- Calculate this amount for the sample `[-6, -2, 1, 7]`

Getting to the variance

Idea 2: average the absolute value of the $x_i - \overline{x}$ difference

$$\sum_{i=1}^{n} \frac{|x_i - \overline{x}|}{n}$$

- Calculate this amount for the sample `[-6, -2, 1, 7]`

$$\frac{|-6-0| + |-2-0| + |1-0| + |7-0|}{4} = 4 \quad \text{☹}$$

Moreover, absolute value is not differentiable at 0

This is inconvenient: `https://www.youtube.com/watch?v=sHRBg6BhKjI`

Getting to the variance

Idea 3: average the squared differences $x_i - \overline{x}$

$$\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}$$

- Calculate this amount for the sample `[-4, -4, 4, 4]`

Getting to the variance

Idea 3: average the squared differences $x_i - \overline{x}$

$$\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}$$

- Calculate this amount for the sample `[-4, -4, 4, 4]`

$$\frac{(-4-0)^2 + (-4-0)^2 + (4-0)^2 + (4-0)^2}{4} = 64 \quad \text{☺}$$

Getting to the variance

Idea 3: average the squared differences $x_i - \overline{x}$

$$\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}$$

- Calculate this amount for the sample `[-6, -2, 1, 7]`

Getting to the variance

Idea 3: average the squared differences $x_i - \overline{x}$

$$\sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n}$$

- Calculate this amount for the sample `[-6, -2, 1, 7]`

$$\frac{(-6-0)^2 + (-2-0)^2 + (1-0)^2 + (7-0)^2}{4} = 90 \quad \text{☺}$$

Source: Example adapted from

`https://www.mathsisfun.com/data/standard-deviation.html`

The larger the better

- The expected value of a (discrete) random variable:

$$E[X] = p(x_1)x_1 + p(x_2)x_2 + \ldots + p(x_n)x_n$$

- Sample mean $\overline{x} \rightarrow$ normalised sum of n i.i.d. random variables

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

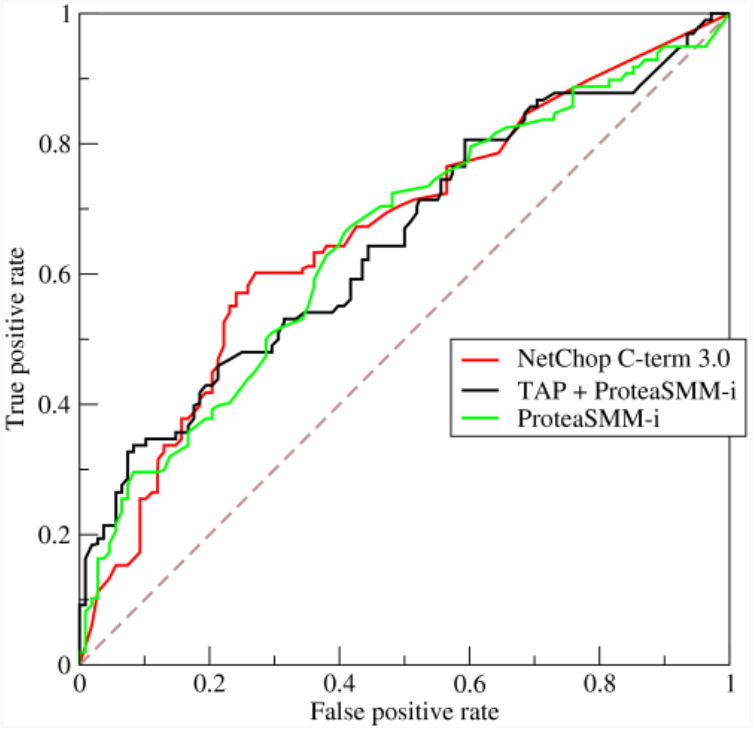
- The law of large numbers states that $\overline{x} \rightarrow E[X]$ for large n

  $\rightarrow$ The (sample) mean $\overline{x}$ is an estimator of the expected value $E[X]$

The larger the sample, the better $\overline{x}$ approximates "true" mean $E[X]$

ROC curve

ROC curves (*Receiver Operating Characteristic*) are very useful to chose a threshold.

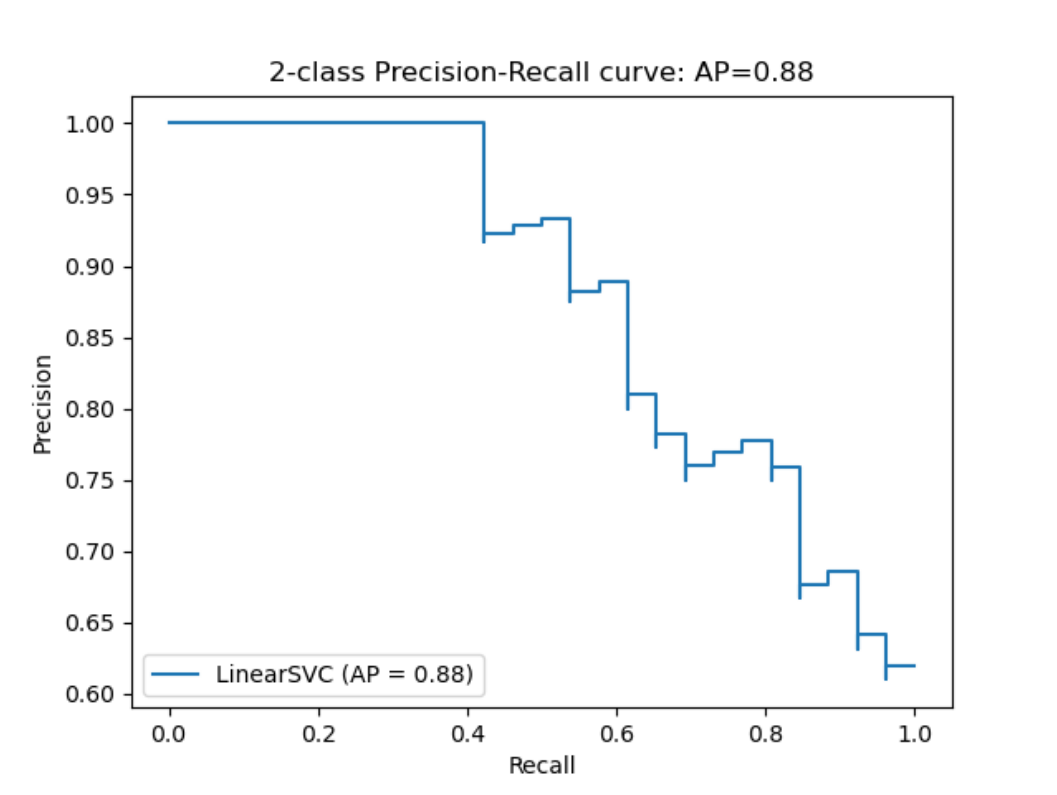


The AUC (*Area Under ROC*) is often used to estimate the model skill.

Image from Wikipedia

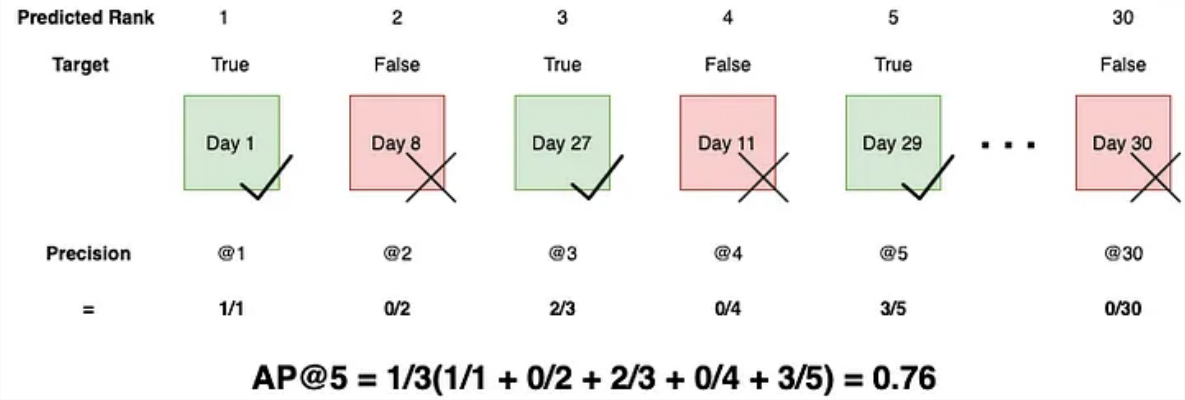
Precision-recall curve

Another way to do this is to use the Precision and the Recall instead of using the True positive and the False positive rates.



Mean average precision

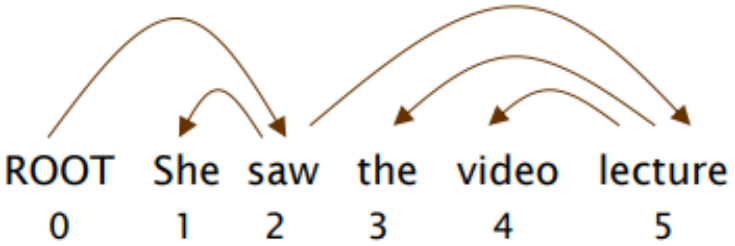
- Model predicts a numerical score
- Gold class is binary or discrete
- Evaluate without setting a fixed threshold



Structured prediction

- How to compare structured objects?
  - $\longrightarrow$ Sub-sequences
  - $\longrightarrow$ Clusters
  - $\longrightarrow$ Syntax trees
  - $\longrightarrow$ Graphs

Structured prediction example: LAS/UAS



$$\text{Acc} = \frac{\text{\# correct deps}}{\text{\# of deps}}$$

"unlabelled attachment score"

UAS = 4 / 5 = 80%

LAS = 2 / 5 = 40%

"labelled AS"

Gold			
1	2	She	nsubj
2	0	saw	root
3	5	the	det
4	5	video	nn
5	2	lecture	obj

Parsed			
1	2	She	nsubj
2	0	saw	root
3	4	the	det
4	5	video	nsubj
5	2	lecture	ccomp

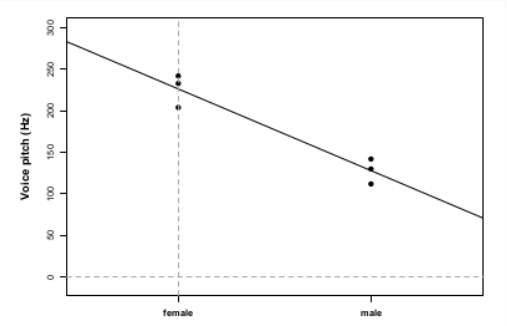
Source: `https://x-wei.github.io/xcs224n-lecture5.html`

Information theory

- Entropy: alternative view of variability/skewness
  - $\rightarrow$ $H = -\sum p(x_i) \log p(x_i)$ $\rightarrow$ amount of uncertainty
  - $\rightarrow$ $H = \max$ for uniform distribution (unpredictable)
  - $\rightarrow$ $H = 0$ for highly skewed distribution (predictable)
- Other useful notions:
  - $\rightarrow$ Cross entropy
  - $\rightarrow$ Mutual information
  - $\rightarrow$ Kullbak-Leibler divergence (asymmetric)
  - $\rightarrow$ Jensen–Shannon divergence (symmetric)

Models for categorical variables

- ANOVA: Generalise t-test for more than 2 means
- Linear model: predict a linear regression slope
  - $\rightarrow$ Is the slope significantly different from zero ?
  - $\rightarrow$ Notation: pitch $\approx$ sex $+\varepsilon$
- Mixed model: more sophisticated for multiple factors



Source: https://bodo-winter.net/tutorials.html

Statistics libraries

- Visual: Excel, Libreoffice, ...
- Python: `matplotlib`, `numpy`, `scipy`, `sklearn`, ...
- R: multiple libraries including linear models
- Proprietary: Matlab, SPSS, ...